Česká zemědělská univerzita v Praze

Provozně ekonomická fakulta

Katedra statistiky

Diplomová práce

Prediktivní modelování odchodu zaměstnanců

Bc. Radka Horníková

ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA V PRAZE

Provozně ekonomická fakulta

ZADÁNÍ DIPLOMOVÉ PRÁCE

Bc. Radka Horníková

Systémové inženýrství a informatika
Informatika

Název práce

Prediktivní modelování odchodu zaměstnanců

Název anglicky

Predictive modeling of employee leave

Cíle práce

Cíl diplomové práce spočívá ve vytvoření modelu predikujícího odchod zaměstnanců z organizace a identifikaci faktorů majících na odchod zaměstnanců největší vliv.

Metodika

Práce bude založena na využití statistických metod a nástrojů data mining k analýze rozsáhlého souboru dat z oblasti lidských zdrojů. Vzhledem k charakteru dat budou aplikovány klasifikační algoritmy, jako jsou rozhodovací strom či náhodný les a logistická regrese. Na základě porovnání predikční schopnosti vybraných modelů bude vytvořen model predikující odchod zaměstnanců z organizace. Rovněž budou určeny faktory nejvíce ovlivňující rozhodnutí zaměstnanců odejít z organizace.

Doporučený rozsah práce

60 – 80 stran

Klíčová slova

Big data, statistická analýza, prediktivní modelování, data mining, lidské zdroje

---

Doporučené zdroje informací

ABBOTT, D. Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst. United States, Indiana: Wiley, 2014. ISBN 978-1-118-72793-5.

ARMSTRONG, M. *Řízení lidských zdrojů : nejnovější trendy a postupy : 10. vydání.* Praha: Grada, 2007. ISBN 978-80-247-1407-3.

AYYADEVARA, V. K. Pro Machine Learning Algorithms: A Hands-On Approach to Implementing Algorithms in Python and R. United States, CA: Apress, 2018. ISBN 978-1-4842-3564-5.

DEAN, J. Big Data, Data Mining and Machine Learning: Value Creation for Business Leaders and Practitioners. New Jersey: Wiley, 2014. ISBN 978-1-118-61804-2.

DIEZ, F., BUSSIN, M., LEE, V. Fundamentals of HR Analytics, UK: Emerald Publishing Limited, 2019. ISBN 978-1-78973-961-9

MELOUN, M. – MILITKÝ, J. – UNIVERZITA KARLOVA. *Kompendium statistického zpracování dat.* Praha: Karolinum, 2012. ISBN 978-80-246-2196-8.

TUFFÉRY, S. Data Mining and Statistics for Decision Making. UK, West Sussex: Wiley, 2011. ISBN 978-0-470-68829-8.

Předběžný termín obhajoby

2020/21 LS – PEF

Vedoucí práce

Ing. Tomáš Hlavsa, Ph.D.

Garantující pracoviště

Katedra statistiky

Elektronicky schváleno dne 23. 11. 2020

prof. Ing. Libuše Svatošová, CSc.

Vedoucí katedry

Elektronicky schváleno dne 24. 11. 2020

Ing. Martin Pelikán, Ph.D.

Děkan

V Praze dne 12. 12. 2020

Čestné prohlášení

Prohlašuji, že svou diplomovou práci "Prediktivní modelování odchodu zaměstnanců" jsem vypracovala samostatně pod vedením vedoucího diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou citovány v práci a uvedeny v seznamu použitých zdrojů na konci práce. Jako autorka uvedené diplomové práce dále prohlašuji, že jsem v souvislosti s jejím vytvořením neporušil autorská práva třetích osob.

V Praze dne 21.3.2021 _____

Poděkování

Ráda bych touto cestou poděkovala Ing. Tomášovi Hlavsovi, Ph.D. za odborné vedení práce, věnovaný čas a cenné rady poskytnuté při konzultacích.

Poděkování patří též mé rodině za podporu během studia.

Prediktivní modelování odchodu zaměstnanců

Abstrakt

Diplomová práce se věnuje prediktivnímu modelování odchodu zaměstnanců a rovněž faktorům ovlivňujícím odchod zaměstnanců.

Teoretická část práce se zabývá nejprve daty obecně a poté se více zaměřuje na big data. Dále se věnuje oblasti datové analýzy a prediktivnímu modelování, přičemž vzhledem k tématu práce se zaměřuje na modely klasifikační, konkrétně na metody Naïve Bayes, logistickou regresi, rozhodovací strom a náhodný les. Teoretickou část uzavírá kapitola věnovaná oblasti lidských zdrojů.

V části praktické jsou poznatky získané v teoretické části práce aplikovány při prediktivním modelování. Postupně jsou na základě dat obsahujících anonymizované záznamy o zaměstnancích vytvořeny modely Naïve Bayes, model logistické regrese, rozhodovací strom a náhodný les. Na základě porovnání kvality vytvořených modelů je vybrán nejvhodnější model pro predikci odchodu zaměstnanců. Pomocí vybraného modelu jsou v závěru práce identifikovány faktory mající na odchod zaměstnanců největší vliv.

Klíčová slova: big data, predikce, prediktivní modelování, CRISP-DM, klasifikace, Naïve Bayes, logistická regrese, rozhodovací strom, náhodný les, lidské zdroje

Predictive modeling of employee leave

Abstract

The diploma thesis focuses on predictive modeling of employee leave and on factors influencing employee leave.

The theoretical part of the thesis first talks about data in general and then focuses more on big data. It also deals with the field of data analysis and predictive modeling, and with regard to the topic of the work focuses on classification models, specifically on the methods of Naïve Bayes, logistic regression, decision tree and random forest. The theoretical part concludes with a chapter devoted to human resources.

In the practical part of the diploma thesis, the knowledge gained in the theoretical part is applied in predictive modeling. Throughout the practical part, Naïve Bayes models, a logistic regression model, a decision tree and a random forest are created based on data containing anonymized employee records. Based on a comparison of the quality of the created models, the most suitable model for predicting of employee leave is selected. Based on a comparison of the quality of the created models, the most suitable model for predicting employees leaving is selected. Using the selected model, the factors that have the greatest influence on the employees leave are identified at the end of the work.

Keywords: big data, prediction, predictive modeling, CRIPS-DM, classification, Naïve Bayes, logistic regression, decision tree, random forest, human resources

Obsah

Seznam obrázků

Seznam tabulek

Seznam grafů

1 Úvod

Odchody zaměstnanců představují pro organizace negativní jev. Současně s odcházejícím zaměstnancem totiž ztrácejí nejen cenné znalosti, jež zaměstnanec během svého působení u organizace nabyl, ale samotný odchod zaměstnance sebou přináší i výrazné náklady. Jedná se nejen o náklady mzdové a administrativní, ale též o náklady spojené se získáváním náhrady za odcházejícího zaměstnance a následné zaučování. Nákladem jsou rovněž i ztráty způsobené dočasnou neobsazeností pracovního místa.

Působením na faktory mající vliv na odchod zaměstnanců, mohou organizace míru odchodů zaměstnanců ovlivnit. Existují faktory, které zaměstnavatel ovlivnit nemůže. Typicky je tomu tak v případě stěhování zaměstnance. Avšak existují faktory, které ovlivnit může.

V současné době, kdy existují výkonné počítače a analytické metody, umožňující zpracovávat rozsáhlá množství dat, dostávají organizace k dispozici nástroje, pomocí nichž mohou odchody zaměstnanců předvídat. Předpokladem je sběr dostatečného množství relevantních a kvalitních dat. Ta je poté možné využít k prediktivnímu modelování, což je oblast analýzy dat, jejímž cílem je předvídání budoucího vývoje sledovaného jevu. V daném případě je sledovaným jevem nastání či nenastání odchodu zaměstnance.

Prvá, teoretická, část diplomové práce, se zaměřuje nejprve na data obecně, poté na big data. Dále se zabývá analýzou dat a prediktivním modelováním. Vzhledem k charakteru problematiky, jež spočívá v modelování odchodu zaměstnanců, je v teoretické části pozornost věnována především modelům klasifikačním a též oblasti lidských zdrojů.

Druhá, praktická, část práce spočívá v samotném prediktivním modelování odchodu zaměstnanců. Nejprve jsou vytvořeny klasifikační modely Naïve Bayes, logistické regrese, rozhodovacího stromu a náhodného lesa. Na základě porovnání jejich kvality je vybrán nevhodnější model pro predikci. Pomocí vybraného modelu jsou poté identifikovány faktory mající na odchod zaměstnanců největší vliv.

2 Cíl práce a metodika

2.1 Cíl práce

Cílem diplomové práce je vytvoření modelu predikujícího odchody zaměstnanců z organizace a identifikaci faktorů majících na odchod zaměstnanců největší vliv. Odchody zaměstnanců se přitom rozumí ukončení pracovního poměru na základě rozhodnutí zaměstnance.

2.2 Metodika

Teoretická část práce vychází z poznatků nabytých na základě literární rešerše. Čerpáno je z odborné literatury české i zahraniční. Využito je rovněž zdrojů elektronických v podobě odborných článků a studií českých i zahraničních, zaměřujících se především na témata big data, prediktivní modelování a oblast lidských zdrojů.

Praktická část práce spočívá v analýze souboru dat obsahující záznamy o zaměstnancích. Data pochází z volně dostupného zdroje, jímž je platforma Kaggle (2020). Data jsou vzhledem ke své citlivé povaze plně anonymizovaná, tzn. neobsahují žádné údaje, které by mohly být byť nepřímo spojeny s konkrétní fyzickou či právnickou osobou.

Data jsou v praktické části analyzována za využití softwarového nástroje RapidMiner (Mierswa, Klinkenberg, 2020). Postup uplatněný při prediktivní analýze vychází z metodiky CRISP-DM, která sestává z šesti fází, kterými jsou porozumění problematice, porozumění datům, příprava dat, modelování, vyhodnocení a využití výsledků.

V části věnované porozumění datům jsou využívány základní popisné statistiky, jakými jsou minimální a maximální hodnoty, nejčastěji a nejméně často vyskytující se hodnoty, průměrné hodnoty, směrodatné odchylky, počty chybějících hodnot a typy proměnných.

Část zabývající se přípravou dat využívá transformací proměnných v případech, kdy je zjištěno, že typ proměnné neodpovídá jejímu obsahu. Výběr vysvětlujících proměnných

neboli prediktorů vhodných pro modelování v softwarovém nástroji RapidMiner spočívá ve vyhodnocení na základě tzv. kritérií CISMT. Kritérium C (correlation) vychází z korelace prediktoru s cílovou proměnnou, kritérium I (ID-ness) vychází z počtu unikátních hodnot prediktoru, kritérium S (stability) naopak z počtu shodných hodnot prediktoru, kritérium (M) z počtu chybějících hodnot daného prediktoru a kritéria T (text-ness) zkoumajícího počet hodnot proměnných sestávajících z více než jednoho slova a rovněž délky hodnot proměnných. Data jsou v rámci přípravy dat též rozdělena v poměru 60:40 na data trénovací a testovací (Ayyadevara, 2018).

V rámci části praktické práce věnované modelování, jsou postupně vytvořeny jednotlivé modely. Vzhledem ke skutečnosti, že cílová proměnná je binární a jsou dostupná historická data, jedná se o úlohu klasifikační a tudíž i využité modely jsou modely klasifikační.

Prvním vytvořeným modelem je naivní Bayesův klasifikátor vycházející z Bayesovy věty o podmíněných pravděpodobnostech. Druhým modelem je model logistické regrese založené na logitové transformaci spočívající v poměru šancí, což je poměr pravděpodobnosti výskytu sledovaného jevu a pravděpodobnosti, že daný jev nenastane. Logitové transformace je docíleno pomocí přirozeného logaritmu šance. Třetím modelem je klasifikační rozhodovací strom, jakožto hierarchické uspořádání rozhodovacích pravidel. Čtvrtým modelem je klasifikační náhodný les. Náhodný les spočívá ve vytvoření vícero modelů stejného typu a následného zkombinování jejich výsledků do finální predikce.

K vyhodnocení kvality modelů je využito matic záměn modelů a též ROC křivek. Matice záměn slouží k porovnání predikovaných hodnot s hodnotami skutečnými. Obsahuje četnosti správně a nesprávně klasifikovaných záznamů. Na základě matice záměn lze získat hodnocení kvality modelu, mimo jiné celkovou správnost, chybu, senzitivitu, specificitu. ROC křivka vychází z matice záměn. Je grafickým zobrazením vztahu mezi senzitivitou a specificitou. Je vhodná k využití při porovnávání kvality více modelů mezi sebou.

3 Teoretická východiska

S rozvojem technologií rapidně narostlo množství generovaných a zaznamenávaných dat prakticky ve všech oblastech vědy a průmyslu. Toto enormní množství dat dává nové možnosti při rozhodování. Pokračující nárůst dat ve světě je poháněn vícero faktory. Patří mezi ně nejen stále narůstající výpočetní výkon a úložná kapacita, či narůstající počet uživatelů internetu a sociálních sítí, ale rovněž samotná hodnota dat, kterou si stále více uvědomují společnosti napříč světem a za účelem sběru dat budují rozsáhlé datové sklady. Tyto datové sklady následně využívají k těžení informací z dat, o které poté opírají svá rozhodnutí. Zpracování takového velkého množství dat však představuje výzvy a vyžaduje inovativní statistické metody.

Dle poznatků Mayer-Schönbergera a Ciglera (2013) je současná společnost přímo zavalena daty, avšak z převážné části je nedokáže zpracovat. Uvádí, že až 85 % dat v Evropě a až 80 % dat ve světě není využíváno. Přitom nevyužitá data sama o sobě nemají význam. Je třeba umět data zpracovat, propojit a vytěžit z nich informace. Hodnota dat je odvozena od schopnosti jejich využití, přičemž za klíčové při jejich zpracování uvedení autoři uvádějí umění přemýšlet o datech a též kladení si správných otázek, na které by data mohla odpovědět. Nevyužíváním dat společnost přichází nejen o jejich hodnotu, ale též o inovace, které jsou pro firmy zásadní. V rozšiřování otevřenosti a sdílení dat v evropském prostoru spatřují šanci Evropy konkurovat digitálním platformám Číny a Spojených států.

3.1 Data

Data, jakožto údaje popisující určitý jev či vlastnost sledovaného objektu, jsou zaznamenávaná zpravidla v digitální podobě a určená k zpracování prostředky výpočetní techniky. V úplném základu se jedná o posloupnost bitů uloženou v paměti. Interpretací těchto dat se z nich stávají informace.

Data jsou získávána měřením či sběrem, shromažďována v úložištích a následně analyzována, přičemž je možné je vizualizovat pomocí grafů, což zpravidla vede k lepšímu porozumění jejich vlastností a vazeb. Na základě analýzy dat je možné vytvářet různé modely, jež mohou následně sloužit kupříkladu k predikci či prognóze možného budoucího vývoje nebo například k nalezení vazeb mezi proměnnými a určení síly těchto vztahů.

3.1.1 Druhy dat

Data je možné na základě úrovně zpracování dělit na data zpracovaná a data nezpracovaná.

Nezpracovaná data (angl. raw data) jsou v takové podobě, v jaké byla sesbírána a shromážděna. Tedy bez jakýchkoli úprav. Nezpracovaná data zpravidla nejsou vhodná k přímé analýze. Je zapotřebí je tzv. očistit, čímž se mají na mysli činnosti, mezi které patří například oprava dat, vhodné naložení s chybějícími hodnotami, převod dat do vhodného formátu, případně řešení výskytu odlehlých či extrémních hodnot. Po očištění se data stávají daty zpracovanými, připravenými k analýze.

Data, jež popisují sledované objekty, lze na základě toho, zda je možné je kvantifikovat či nikoli, dělit na data kvantitativní (též numerická) a data kvalitativní (též kategoriální).

Data kvantitativní jsou reprezentována číselnými hodnotami. Lze je dále dělit na data spojitá a diskrétní. Spojitá data mohou nabývat libovolných hodnot v určitém intervalu (například výška či váha). Data diskrétní nabývají pouze spočetně mnoha hodnot. Na reálné ose jsou reprezentována izolovanými body (například počet dětí).

Data kvalitativní jsou naopak data, která nelze kvantifikovat, ale lze je roztřídit do kategorií. Příkladem může být například dosažené vzdělání či barva vlasů. Je sice možné, aby jednotlivé kategorie byly reprezentovány číselnými hodnotami, avšak tyto číselné hodnoty označující kategorie nemají v takovém případě souvislost s úrovní sledovaného znaku.

Kvalitativní data je dále možné dělit dle toho, zda je možné je seřadit dle určitého hlediska a přiřadit jim pořadí. Na základě této skutečnosti dělíme data kvalitativní na data ordinální, nominální a binární.

Ordinální data obsahují více kategorií, jež je možné seřadit (například dosažené vzdělání). Nominální data taktéž obsahují více kategorií, avšak seřadit je nelze (například barva vlasů).

Binární data jsou taková data, která mohou nabývat pouze dvou hodnot, které obecně vyjadřují, zda se sledovaný jev vyskytuje či nevyskytuje. Typicky jsou vyjádřena hodnotami 0 a 1.

3.1.2 Big Data

Samotný pojem big data napovídá, že se jedná o data obsáhlá. Avšak neexistuje přesná hranice rozdělující data na velká a malá. Označení big data obecně popisuje taková data, jež člověk ani technika nejsou schopni běžně dostupnými způsoby zpracovat.

Počátek zavedení pojmu big data lze vysledovat do roku 1997, kdy Cox a Ellsworth (1997) tento pojem použili ve své práci, přičemž big data definovali pomocí popisu stavu, ve kterém data dosáhnou takového objemu, že je již není možné uchovávat v hlavní paměti počítače a zároveň ani na lokální disk počítače.

Odborníci se však snažili o přesnější popis big data, přičemž k němu přistoupili pomocí popisu tří charakteristických vlastností big data. Tyto charakteristiky popsal Laney (2001) jako tzv. 3V, odvozená z anglických slov volume, velocity a variety, které v českém překladu znamenají velikost, rychlost nárůstu a různorodost.

Rovněž Beyer a Laney (2012) popsali big data jako data, jejichž velikost (angl. volume), rychlost nárůstu (angl. velocity) a různorodost (angl. variety) neumožňují jejich zpracování v rozumném čase prostřednictvím do té doby běžně využívaných technologií.

V průběhu následujících let se k třem charakteristickým vlastnostem big data přidaly další. Dle Normandeau (2013) mezi ně patří nejistá věrohodnost (angl. veracity) dat, dále limitovaná doba platnosti (angl. validity) dat pro jejich využití a s tím související přechodná doba jejich nutného ukládání (angl. volatility). Firican (2017) uvádí vcelku deset V, přičemž k dříve uvedeným přidává zranitelnost (angl. vulnerability) dat, vyzdvihující možné riziko spojené s únikem dat z rozsáhlých datových souborů, a dále vizualizaci (angl. visualization) dat a vysokou hodnotu (angl. value), kterou big data mají pro společnost či organizaci, jež je vlastní.

Dean (2014) big data charakterizuje obecně jako data, která nemohou být zpracována bez speciálního softwaru. V případě enormně rozsáhlých dat jsou využívány distribuované systémy založené na rozložení výpočetních úloh mez vícero počítačů nebo serverů vzájemně propojených sítí. Tato technologie se vyvinula v důsledku příliš vysoké ceny tzv. superpočítačů. Využitím technologií distribuovaných systémů mohou být zpracovány velmi rozsáhlé objemy dat s vysokým počtem atributů.

Dean (2014) dále charakterizuje big data prostřednictvím jejich nároků na uložení, zpracování či analyzování, které přesahují z důvodu svého objemu možnosti tradičních postupů. Toto může zahrnovat nejen výpočetní výkon, ale též i nedostatek času k zpracování dat a získání informace k učinění rozhodnutí tradičními metodami.

Dean (2014) též jako typické pro big data uvádí jejich využívání v maximální míře, tj. využívání veškerých dostupných dat, nikoli pouze jejich části. Důvodem je především skutečnost, že je-li k dispozici více dat, obzvláště pak dat historických, zpravidla to vede k lepší, přesnější predikci. Rovněž využívání relativně nových zdrojů, jakými jsou především sociální sítě, pomáhá lepším predikcím týkajícím se zájmů či potřeb zákazníků a jejich preferencí.

K porovnání metod využívaných při práci s big data s klasickými statistickými přístupy, založenými na vzorkování, Dean (2014) uvádí, že pro mnoho úloh klasické statistické metody přinesou stejně dobré výsledky jako analyzování celé populace, tj. všech dostupných dat. Nicméně existuje i značná část úloh, především těch založených na nalézání odchylek, kde toto vždy neplatí. Často v případě, kdy by určitá organizace vycházela pouze ze vzorku celé populace, mohla by přijít o důležité poznatky. Z pohledu statistiky má odlehlá hodnota zpravidla negativní význam. Avšak v mnoha případech mohou odlehlé hodnoty naopak představovat nejpřínosnější skupinu zákazníků či novou cílovou skupinu zákazníků, která může být na základě své odlišnosti zkoumána a efektivně využita. U těchto typů úloh je nezbytné využívat všechna dostupná data, nikoli pouze jejich část. Příkladem může být například kontrola kvality ve výrobě. Dalším příkladem je detekce podvodů, kdy jsou odhalovány podvody například s kreditními kartami. Předpokládejme, že 1 z 1000 transakcí je podvodná. Pokud bychom použili pouze 20 % vzorek dat, mohlo by se stát, že bychom neměli jediný záznam podvodné transakce ve vzniklé podmnožině původních dat. Přitom ale například pro prediktivní modelování, které je založené na využití údajů z historie k vytvoření modelu rozeznávajícího podvod, je zcela nezbytné zachytit při vytváření modelu pokud možno maximum těchto podvodných operací, aby byl model efektivní pro predikci budoucích podvodných transakcí. Pokud by model vycházel pouze ze vzorku a nikoli z veškerých dostupných dat, hrozilo by, že by model neodhalil ani budoucí podvodné aktivity.

Samotnému efektivnímu využití big data předchází jejich vhodné uložení a zpracování. V případě, jedná-li se o druh dat, která neustále přibývají, tj. přichází v proudu (angl. stream), je nezbytné tato data rovněž průběžně i zpracovávat, a to navíc dostatečně rychle. Tento typ úloh je nazýván termínem OLTP (angl. online transaction processing) a zpravidla je obstarává databázový systém, který v reálném čase data aktualizuje a zároveň přímo i zpracovává (Holubová et al., 2015).

Inmon (2002) jako možné řešení uložení big data uvádí tzv. datový sklad, jenž definuje jako kolekci sjednocených a předmětově orientovaných databází umožňujících přístupu

k informacím sloužícím jako podklad pro rozhodování. V případě datových skladů se předpokládá uložení velkých objemů dat a provádění rozsáhlých analytických operací s daty. Berka (2003) jako charakteristický znak odlišující datové sklady od databází uvádí orientaci datových skladů na subjekty, které jsou sledovány. Těmito subjekty můžou být například zákazníci, dodavatelé či produkty. Jako další odlišující znak uvádí skutečnost, že datové sklady obvykle uchovávají pouze data, která jsou pro rozhodování relevantní.

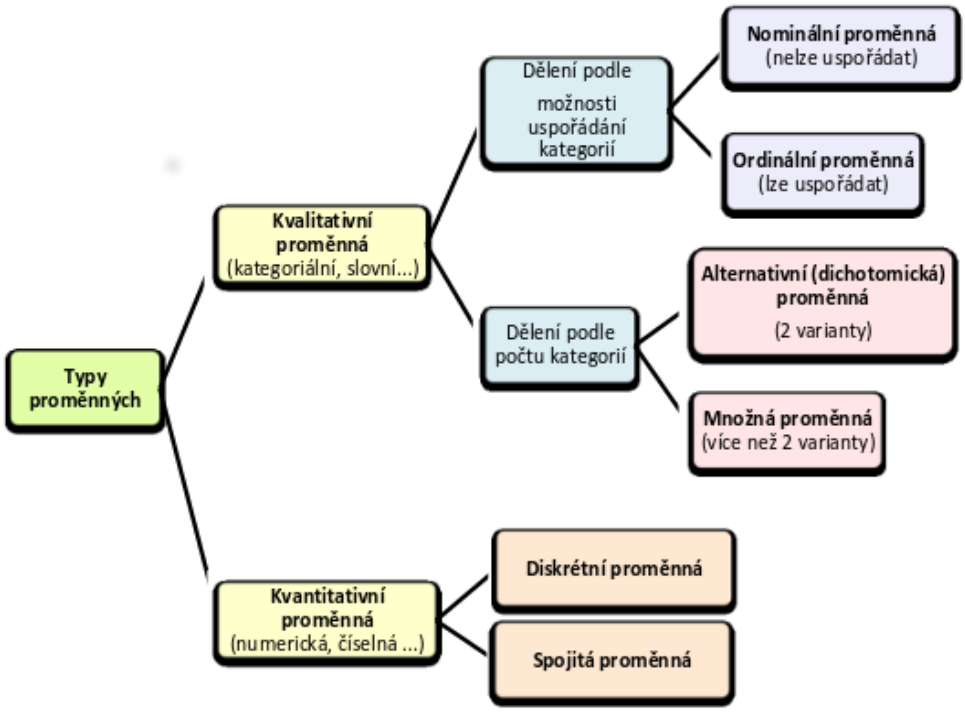
Holubová et al. (2015) proces, pomocí něhož jsou naplňovány datové sklady, popisují jako proces nazývaný pojem ETL. Termín je odvozen od názvů fází procesu v anglickém jazyce. Těmito fázemi jsou extrakce (angl. extract), transformace (angl. transform) a načtení (angl. load). V průběhu extrakce jsou získávána data, často z vícero zdrojů. Během transformace jsou data očištěna od nevyhovujících hodnot a sjednocena. Cílem fáze načtení je uložení dat v datovém skladu.

3.1.3 Typy proměnných

Pomocí proměnných jsou popisovány určité vlastnosti či stavy sledovaných objektů či procesů. Proměnná může například vypovídat o barvě vlasů, úrovni vzdělání anebo výši mzdy.

Proměnné lze v základu rozdělit na dva základní typy. Prvním typem jsou proměnné kvantitativní (též numerické), které nabývají číselných hodnot. Druhým typem jsou proměnné kvalitativní (též nazývané kategoriální), které je možné zařadit do tříd resp. kategorií.

Obrázek 1: Typy proměnných



Zdroj: Litschmannová (2011)

Litschmannová (2011) uvádí, že kvalitativní proměnné je možné na základě možnosti uspořádání dále rozdělit na ordinální (též pořadové) a nominální. Ordinální proměnné je možné určitým způsobem seřadit. Příkladem kvalitativní ordinální proměnné může být například výše mzdy, nabývající hodnot nízká, střední, vysoká. Nominální proměnné seřadit nelze, o jejich hodnotách lze pouze říci, zda se liší, či zda jsou shodné. Příkladem kvalitativní nominální proměnné můžou být například barvy.

Dle jiného hlediska, kterým je dělení podle počtu kategorií, lze kvalitativní proměnné dělit na alternativní a množné. Alternativní proměnné jsou též nazývány jako dichotomické nebo binární. Jsou to proměnné, jež mohou nabývat pouze dvou hodnot. Typickým příkladem je proměnná nabývající pouze hodnot 0 či 1. Dalším příkladem je například proměnná nabývající hodnot muž a žena. Množná proměnná je taková kvalitativní proměnná, která může nabývat více než dvou hodnot. Příkladem množné kvalitativní proměnné je například barva nebo fakulta.

Kvantitativní proměnné je dále možné rozdělit na spojité (též kontinuální) a diskrétní. Diskrétní kvantitativní proměnné, jsou takové proměnné, které mohou nabývat pouze hodnot celočíselných.

Spojité kvantitativní proměnné mohou nabývat libovolných číselných hodnot a lze je dále rozdělit na proměnné intervalové a poměrové (též podílové). Intervalové jsou takové spojité proměnné, u nichž je možné říci, o kolik se jedna hodnota liší od druhé, ale nemá význam určovat jejich podíl. Příkladem může být teplota na Celsiově stupnici. Poměrové spojité proměnné jsou takové proměnné, u nichž lze vypočítat, kolikrát je jedna hodnota větší či menší než druhá hodnota, přičemž podíl hodnot má reálný význam. Příkladem poměrové spojité proměnné je například hmotnost či výška.

3.2 Analýza dat

Dle Kotu (2019) jsou za účelem získávání hodnotných a relevantních informací z rozsáhlých souborů dat kombinovány metody z různých oblastí. Konkrétně uvádí především oblasti statistiky, strojového učení a teorii databází.

Kotu (2019) jako příklady metod analyzování dat uvádí klasifikaci, regresi, asociační analýzu, shlukování, detekci anomálií či odlehlých hodnot, generátory doporučení, výběr prediktorů, prognózy časových řad, hluboké učení neboli deep learning a například též dolování z textu.

Metody supervizované a nesupervizované

Metody využívané k analýze dat Kotu (2019) obecně dělí na supervizované a nesupervizované resp. na tzv. metody učení s učitelem a metody učení bez učitele. Uvádí též, že kompletní aplikace analýzy dat může zahrnovat zároveň supervizované i nesupervizované metody. Nesupervizované metody zpravidla pomáhají porozumět datům. Např. v marketingu metoda tzv. shlukování může být využita k nalezení přirozených shluků v rámci záznamů o zákaznících. Každý zákazník je přiřazen do určitého shluku neboli skupiny, které jsou během procesu vytvořeny na základě podobností

v datech. Rozdělení do shluků může být následně využito pro přiřazování nových zákazníků do kategorií pomocí supervizovaných klasifikačních metod.

Metody supervizované (angl. supervised learning) jsou široce využívané při prediktivním modelování. Pro tyto metody je typická existence cílové proměnné. V rámci supervizovaného učení je hledán funkční vztah mezi cílovou proměnnou a proměnnými nezávislými, které se též nazývají prediktory. Volba konkrétní metody se zpravidla odvíjí od typu cílové proměnné, tj. od oboru hodnot, kterých může tato cílová proměnná nabývat.

V případě, kdy cílová proměnná nabývá několika málo hodnot, jedná se nejčastěji o úlohu tzv. klasifikace (angl. classification). Často cílová proměnná nabývá pouze dvou hodnot, typicky 0 a 1. V takovém případě ji označujeme jako binární. Příkladem takové úlohy je například predikce, zda určitý klient úvěr splatí či nikoli.

V případě, kdy cílová proměnná může nabývat vícero hodnot, tj. je možné ji uvažovat jako spojitou, jedná se zpravidla o úlohu tzv. regrese (angl. regression).

V obou případech, klasifikace i regrese, je vysoce využívaným modelem tzv. rozhodovací strom (angl. decision tree), který lze použít pro oba uvedené typy úloh. Dle typu úlohy je poté rozhodovací strom označován jako strom klasifikační nebo jako strom regresní.

Klasifikační a regresní metody predikují cílovou proměnou na základě nezávislých proměnných (prediktorů). Predikce je založena na zobecněném modelu vytvořeném na základě dříve známých dat. V případě regresních metod je cílová proměnná numerická. Klasifikační úlohy predikují cílovou proměnnou, která je kategorická (též polynomiální).

Klasifikace

Cílem klasifikace je predikce, zda určitý objekt náleží do některé z předdefinovaných skupin. Predikce je založena na předchozím učení z trénovacích, tj. známých, dat. Jedná se o metodu supervizovanou neboli metodu učení s učitelem. Cílová proměnná je kategorická. Využívanými algoritmy jsou především rozhodovací stromy. Dále také neuronové sítě, Bayesovy modely či algoritmus k-nejbližších sousedů (neboli k-NN).

Klasifikace je využívána například v oblasti marketingu, kdy jsou zákazníci na základě hodnot prediktorů zařazování do předem známých skupin (též kategorií).

Regrese

V případě regrese je cílová proměnná spojitá resp. taková, že může nabývat vícero hodnot a je ji možné jako spojitou uvažovat. Stejně jako v případě klasifikace, je i v případě regrese na základě předem známých, tzv. trénovacích, dat vytvořen model, který je následně aplikován na data nová. Mezi algoritmy využívané při regresní analýze patří především lineární regrese a logistická regrese. Příkladem využití regrese je například predikce míry nezaměstnanosti.

3.2.1 Prediktivní modelování

Prediktivní modelování pomáhá vylepšovat obchodní i výrobní procesy napříč obory. Dokáže poskytovat základ pro posouzení různých rizik či příležitostí a je podkladem pro efektivní rozhodování. Proces prediktivního modelování může být zautomatizován a samotný model se dokáže zlepšovat pomocí učení na přibývajících datech.

Roebuck (2011) prediktivní modelování popisuje jako takovou oblast analýzy dat, která má za cíl získávat informace, jež mohou být následně využity k předvídání budoucího pravděpodobného vývoje sledovaného jevu či objektu. Prediktivní analýza kombinuje vícero metod, přičemž využívá jak metody statistické, tak metody data miningu a strojového učení.

Prediktivní analýza se se statistikou značně překrývá. Prediktivní analýza často využívá postupy a testy běžně využívané ve statistice. Přesto existují významné rozdíly mezi těmito dvěma disciplínami, mezi které Abbott (2014) uvádí například rozdíl v datech, kdy ve statistice jsou data typicky menšího rozsahu než v případě prediktivní analýzy, která zpravidla pracuje s daty, jež lze většinou označovat jako big data. S tím souvisí i rozdíl v obou přístupech, a to takový, že postupy prediktivní analýzy nemusí být příliš efektivní v případě malého množství dat a svou sílu předvedou lépe na rozsáhlých souborech dat.

Jako další rozdíl Abbott (2014) uvádí samotný přístup k analýze, ke kterému uvádí, že v případě statistické analýzy metody zpravidla vycházejí z hypotéz, které mají být buďto vyvráceny nebo zamítnuty a rovněž jsou založené na tom, zda data mají či nemají určité očekávané vlastnosti. Statistická analýza též klade důraz na ošetření dat ve smyslu zamezení zkreslení výsledků přítomností odlehlých hodnot. V rámci statistické analýzy je rovněž pozornost věnována zkoumání reziduí z hlediska ověření, zda se data blíží či vzdalují normálnímu rozdělení. Avšak požadavek na normalitu je s rostoucí velikostí dat méně významný, tudíž v případě prediktivního modelování nemá stejný význam jako v případě statistické analýzy. Abbott (2014) též uvádí, že v případě prediktivního modelování se analytici většinou o něco méně zaměřují na samotné hodnoty koeficientů a naopak více se zaměřují na predikční schopnost modelu, neboli jeho přesnost.

Prediktivní analýza má velmi mnoho společného s data miningem. Abbott (2014) k srovnání prediktivního modelování a data miningu uvádí, že do doby, než se prediktivní analýza stala pojmem, považoval uvedené disciplíny za shodné, jelikož algoritmy a přístupy jsou v obou případech vesměs stejné.

Larose (2015) data mining popisuje jako proces nalézání relevantních vztahů a vzorů v rozsáhlých datových souborech. Prediktivní analýzu popisuje jako proces získávání informací z rozsáhlých datových souborů za účelem vytvoření predikce a odhadu budoucího vývoje.

Dle Abbotta (2014) vznikla prediktivní analýza odvozením z vícero příbuzných vědních disciplín, mezi kterými uvádí statistiku, data mining, strojové učení a též i umělou inteligenci. Jako hlavní cíl prediktivní analýzy uvádí hledání závislostí v datech, na základě kterých by bylo možné co nejlépe předpovědět budoucí stav či chování zkoumaného objektu. Analýza je založena na zkoumání vztahu nezávislých proměnných (prediktorů) a cílové proměnné. Součástí procesu prediktivní analýzy je i výběr nejvhodnějších prediktorů a rovněž průběžné vyhodnocování predikční schopnosti vzniklých modelů. Na základě získaného, nejlépe predikujícího modelu, je poté předvídán pravděpodobný budoucí vývoj stavu či chování určitého objektu či procesu. Abbott (2014) zároveň dodává,

že je také možné analýzy provádět za účelem porozumění chování určitého procesu nebo skupiny lidí.

Abbott (2014) upozorňuje na problémové oblasti, na které je třeba pamatovat při prediktivním modelování. V prvé řadě mezi ně patří nevyhovující data, tj. data která jsou buďto nevhodného formátu nebo jsou nevhodná z hlediska obsahu. Základním požadavkem na strukturu dat je dvojrozměrný formát sestávající z jednotlivých sloupců (též atributů), představujících prediktory, a řádků (též záznamů) stejné velikosti. V případě klasifikace je nezbytným předpokladem existence záznamů v datech, představujících výskyty všech možných hodnot cílové proměnné. V opačném případě by se prediktivní model neměl jak naučit predikovat výskyt dané hodnoty cílové proměnné. Jako zásadní problém Abbott (2014) uvádí tzv. přeučení, což je stav, kdy je prediktivní model natolik svázán s trénovacími daty, že se jedná už spíše o popisný model než prediktivní model. Přeučení se následně projevuje velice dobrými výsledky predikce na trénovacích datech, avšak podstatně horšími výsledky na datech testovacích.

Prediktivní modely jsou využívány napříč obory. Například úvěrové společnosti je využívají pro vyhodnocení dlužníka, přičemž využívají dostupné informace o dané osobě, mezi které patří výše příjmů, zůstatek úvěru apod. Své široké uplatnění prediktivní modely nacházejí též v oblasti správy a údržby různých přístrojů a zařízení. Dále v oblasti marketingové, kde se uplatňuje při předvídání chování zákazníka, jehož chování je analyzováno a na základě podobností v chování skupin zákazníků fungují různé doporučující systémy, nabízející zákazníkovi zboží, u kterého je nejvyšší pravděpodobnost, že jej zaujme.

Prediktivní modelování nalézá své uplatnění též v případech zkoumání odchodů zákazníků či zaměstnanců, kdy je možné na základě záznamů o objektech sestavit predikční model popisující typické chování či atributy daného objektu a schopný vyhodnocení nových objektů – predikce zda u daného objektu odchod nastane či nenastane. Dalším oborem, ve kterém prediktivní modelování nalézá široké uplatnění je medicína. Lze například na základě analýzy příznaků vyskytujících se u pacientů vyhodnocovat pravděpodobnou

diagnózu u pacientů nových s podobnými příznaky, a to navíc s určitým předstihem. V oblasti informačních technologií nalézá prediktivní analýza též své uplatnění. Například pomocí prediktivních modelů je odhalován spam mezi emailovými zprávami.

Metody využívané v prediktivní analýze je možné rozdělit na supervizované a nesupervizované. Nejvíce využívanými jsou supervizované metody neboli metody učení s učitelem, spočívající v nalézání vztahů a závislostí mezi prediktory a cílovou proměnnou prostřednictvím vytvoření predikčního modelu na základě známých, tzv. trénovacích, dat a následné aplikaci vytvořeného modelu na nová data.

Často využívaným typem supervizovaných metod v prediktivním modelování jsou metody určené k tzv. klasifikaci, kdy cílová proměnná je kategoriálního typu a účelem je zařazení objektů do předem známých skupin neboli kategorií. Metodami aplikovanými na klasifikační úlohy jsou například logistická regrese, rozhodovací strom, náhodný les, Naïve Bayes.

Dle Roebucka (2011) mezi často využívané modely v prediktivní analýze patří modely regresní. Jako nejvíce využívané uvádí model lineární regrese a model logistické regrese. Model lineární regrese popisuje vztahy mezi nezávislými proměnnými (prediktory) a cílovou proměnnou, která je kvantitativní. V případě modelu logistické regrese je cílová proměnná kvalitativní, kategorická, velmi často je binární, tzn. v takovém případě nabývá právě dvou hodnot.

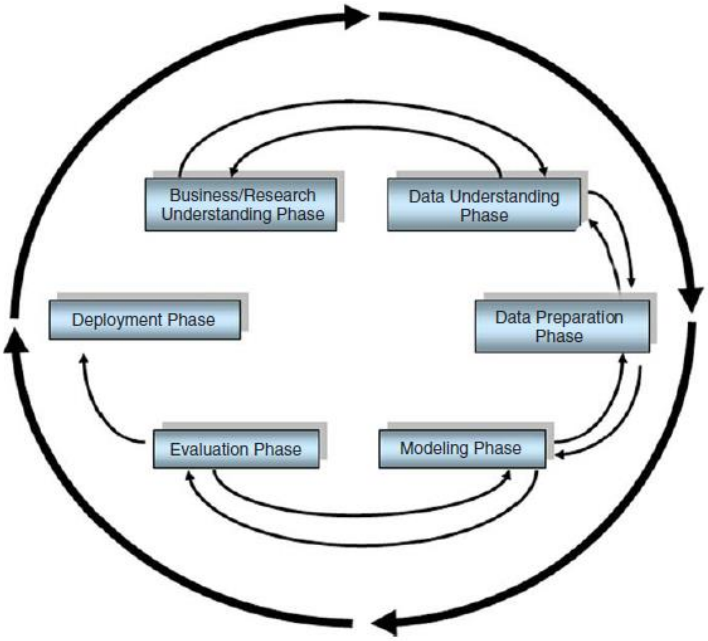
3.2.2 Postup při analýze

Dle Larose (2015) je jednou z nejpopulárnější metodik využívanou při prediktivní analýze metodika CRISP-DM. Sestává z šesti fází, přičemž jejich pořadí se zpravidla přizpůsobuje danému konkrétnímu procesu. Často jsou výstupem jedné fáze nové otázky či problémy, k jejichž řešení je zapotřebí vrátit se do některé z fází předchozích a proces opakovat s využitím nových poznatků.

Dle Larose (2015) metodika CRISP-DM sestává z šesti fází:

1. porozumění problematice,
2. porozumění datům,
3. příprava dat,
4. modelování,
5. vyhodnocení výsledků,
6. využití výsledků.

Obrázek 2: Metodika CRISP-DM



zdroj: Larose (2015)

V prvé fázi, kterou je porozumění problematice, je popsán řešený problém a též je stanoven cíl či cíle prediktivní analýzy.

Druhá fáze, porozumění datům, spočívá ve shromáždění dat a seznámení se s daty prostřednictvím zkoumání jejich základních charakteristik za účelem nalezení možných problémů, které by mohly nastat během prediktivního modelování. Výstupem této fáze je vyhodnocení kvality dat a stanovení požadavků na jejich úpravu v nadcházející fázi.

Třetí fáze, příprava dat, je nejnáročnější fází. Z tohoto důvodu je problematika spadající do této fáze více rozvedena v následující kapitole. Během fáze přípravy dat jsou vybírány proměnné vhodné k analýze. V případě potřeby je prováděna transformace proměnných. Dále je prováděno čištění dat spočívající například v ošetření chybějících hodnot.

Čtvrtá fáze sestává z vytvoření prediktivního modelu či modelů a hodnocení jejich prediktivní síly. V případě potřeby též probíhá optimalizace modelů. Výstupem této fáze mohou být též zjištění vyžadující opětovnou úpravu či transformaci dat, na základě kterých je potřeba vrátit se do třetí fáze, přípravy dat.

Pátou fází je vyhodnocení výsledků, během které je porovnávána predikční schopnost vytvořených modelů a vybrán nejvhodnější model. Rovněž je posouzeno, zda model splňuje cíl stanovený v prvé fázi.

Šestou, poslední fází je fáze využití, kdy je model nasazen do praxe například u zákazníka.

3.2.3 Příprava dat

Příprava dat, jakožto nejnáročnější fáze procesu prediktivní analýzy dle metodiky CRISP-DM, spočívá především v čištění dat, výběru vhodných proměnných a jejich transformaci. Larose (2015) k této problematice uvádí, že mnoho datových souborů uložených v databázích jsou v nezpracované podobě. Mohou být v nevhodném formátu. Často se též stává, že data jsou neúplná. Nebo data mohou obsahovat záznamy, které již nejsou aktuální. Mnohdy se v datech můžou vyskytovat i tzv. redundantní záznamy. Z uvedených důvodů je nezbytné, aby získaná data prošla fází přípravy, sestávající především z čištění dat a transformace dat. V závislosti na konkrétních datech může samotná příprava dat představovat 10 – 60 % času z celého procesy analýzy. Mezi klíčové části přípravy dat v případě klasifikace big data řadí mimo jiné ošetření chybějících hodnot, transformaci kategorických proměnných, odstranění nevýznamných proměnných a binning dat.

Ošetření chybějících hodnot

Chybějící hodnoty jsou častým problémem v datech. Často využívaným postupem je jednoduché vynechání záznamu či záznamů obsahujících chybějící hodnotu. Larose (2015) však upozorňuje, že tento přístup se nemusí vyplatit. V případě, kdy by takovýchto chybějících záznamů, jež budou vymazány, bylo více než velmi malé množství, hrozí, že z původních dat zbyde podmnožina dat, která bude určitým způsobem zkreslená a tedy již nebude odpovídat původním datům. Vynecháním záznamů s chybějící hodnotou může tudíž dojít ke ztrátě cenné informace. Uvedené hrozí i v případě relativně malého počtu vynechaných záznamů. Například v případě detekce podvodů, kterými se zabývají pojišťovny, banky a jiné finanční instituce, kdy by v datech připadala například jedna podvodná transakce na jeden tisíc transakcí, může vynecháním právě tohoto jednoho záznamu dojít ke ztrátě zásadní informace.

Larose (2015) jako vhodnější řešení chybějící hodnoty doporučuje její nahrazení. Toto nahrazení může být provedeno například pomocí konstanty, což ovšem vyžaduje specialistu s velmi dobrou znalostí daných dat. Zpravidla je spíše využíváno nahrazení průměrnou hodnotou v případě proměnných číselných či modem v případě proměnných kategorických. Další možností je nahrazení chybějící hodnoty náhodně vygenerovanou hodnotou pocházející z rozdělení dané proměnné. Též je možné nahrazovat chybějící hodnoty pomocí tzv. imputace dat na základě hodnot ostatních proměnných daného záznamu. Imputace dat usiluje o nalezení nejpravděpodobnější hodnoty v porovnání s ostatními hodnotami daného záznamu.

Larose (2015) však k prvým třem zmiňovaným metodám, tj. nahrazení konstantou, nahrazení průměrem resp. modem a nahrazení náhodně vygenerovanou hodnotou, dodává, že sebou přinášejí problémy v podobě možného zkreslování výsledků. Proto jako nejvhodnější způsob nahrazování chybějících hodnot doporučuje imputaci, tj. nahrazení hodnotou nejpravděpodobnější ve vztahu k ostatním hodnotám daného záznamu. Existují metody imputace chybějících hodnot pro proměnné spojité a metody pro imputace hodnot pro proměnné kategoriální. V případě hodnot spojitých lze například pomocí regresního

modelu odhadnout chybějící hodnotu, přičemž výchozím souborem dat pro regresi by byly všechny ostatní záznamy v datech, tj. data bez chybějících hodnot. V případě proměnné kategorické lze podobně využít například pravidel vytvořených pomocí rozhodovacího stromu, sestaveného z dat neobsahujících chybějící hodnoty.

Nalezení odlehlých hodnot

Larose (2015) k odlehlým či extrémním hodnotám uvádí, že někdy mohou pomoci odhalit chyby v datech. Někdy však odlehlá hodnota může přestavovat i zcela validní data, nikoli chybu. Některé, především statistické metody, mohou být na odlehlé hodnoty citlivé a proto je v takovém případě nežádoucí, aby byly v datech přítomné, přestože se jedná o hodnoty validní.

Transformace kategorických proměnných do dummy proměnných

Larose (2015) upozorňuje na skutečnost, že některé analytické metody, jako například regrese, vyžadují, aby byly prediktory v číselné podobě. Proto v takovém případě, pokud data obsahují proměnné kategorické, je zapotřebí je transformovat pomocí tzv. dummy proměnných. Zpravidla, pokud má proměnná *k* kategorií, je zapotřebí $k - 1$ dummy proměnných. Pro tzv. referenční kategorii není totiž třeba dummy proměnnou vytvářet, jelikož je ji možné vyjádřit pomocí hodnot ostatních dummy proměnných.

Transformace kategorických proměnných do číselných proměnných

Larose (2015) upozorňuje na nebezpečnou chybu spočívající v transformování kategorických proměnných do číselných proměnných v případě, kdy původní kategorické proměnné nejsou ordinální, tj. není možné je seřadit (např. barvy nebo světové strany). V takovém případě varuje před transformací dat do formátu typu: 1, 2, 3, 4. A to z důvodu možného zkreslení vztahu mezi hodnotami, kdy by se z neporovnatelných hodnot staly porovnatelné. V případě kategorických proměnných ordinálních připouští možnost transformace do číselných proměnných, ale obecně tento postup spíš též nedoporučuje, jelikož by mohlo též dojít ke zkreslení spočívajícím v možném skutečnosti neodpovídajícím vzájemném poměru hodnot.

Binning dat

Larose (2015) binning dat, též nazývaný jako diskrétní binování či diskretizace, popisuje jako metodu přípravy dat, využívanou k snížení chyby modelu. Spočívá v nahrazení určitého intervalu původních dat, tzv. binu, hodnotou představující tento interval, zpravidla střední hodnotou. Některé algoritmy preferují kategorické proměnné před spojitými. V takovém případě je možné rozdělit číselné prediktory do jednotlivých binů a nahradit je hodnotami reprezentujícími tyto jednotlivé biny. Například v případě prediktoru, jenž by popisoval ceny zboží, je možné biny číselných hodnot nahradit například hodnotami typu: nízké, střední, vysoké.

Ekvivalentem binningu číselných hodnot pro kategorické proměnné je jejich reklasifikace. Často kategorické proměnné obsahují mnoho různých hodnot, například proměnná stát nebo proměnná město. Metody, jako například logistická regrese nebo rozhodovací strom C4.5, nepracují nejlépe, pokud mají zpracovávat proměnné kategorické, které obsahují příliš mnoho různých hodnot. Larose (2015) doporučuje v takovém případě reklasifikaci kategorické proměnné. Například v případě velkého počtu různých hodnot proměnné představující města, doporučuje reklasifikovat například na regiony a snížit tak počet různých hodnot, kterých daná proměnná může nabývat.

Odstranění nevýznamných proměnných

Proměnné, které nepřispívají k vylepšení výsledků analýzy a zároveň neobsahují nezbytnou informaci pro modelování Larose (2015) doporučuje odstranit. Vyjmenovává především proměnné unární nebo téměř unární, což jsou takové proměnné, které jsou představovány jednou hodnotu, tj. jedná se v podstatě o konstantu. Uvádí, že i v případě, kdy je proměnná téměř unární, tj. téměř konstantou, je zpravidla vhodnější ji nezahrnovat. Může se ovšem stát, že i taková proměnná bude pro model významná a doporučuje před jejím odstraněním ověřit její vliv na výsledky modelování.

3.2.4 Klasifikace

Dle Kotu (2019) je klasifikace nejvíce využívanou skupinou metod při analýze dat. Principem klasifikace je využitím klasifikačního modelu predikovat cílovou proměnnou, která je binární či kategorická na základě hodnot prediktorů. Model nalézá zobecněný vztah mezi proměnnými na základě trénovacích dat.

Holčík a Komenda (2015) klasifikaci dat popisují jako oblast analýzy dat, jež spočívá v rozdělení sledovaných subjektů do předem definovaných skupin. Uvádí, že v počáteční fázi klasifikace jsou obvykle za účelem lepšího výsledku klasifikace data redukována, tzn. jsou vyjádřena v menším počtu proměnných, které od sebe lépe odliší jednotlivé skupiny objektů.

Klasifikace je využívána jednak v případě řešení tzv. rozpoznávacích úloh, kdy je potřeba rozhodnout o typu či charakteru objektu nebo kdy je posuzována kvalita stavu určitého objektu. Rovněž klasifikace je využívána v případě řešení tzv. predikčních úloh, tj. při rozhodování o budoucím vývoji sledovaného objektu. Existují i další metody, jež z hlediska metodologie spadají pod klasifikaci dat, jelikož jsou na ní založené. Například v případě klasifikace dat, kdy skupiny by nebyly předem definovány, by byla použita metoda tzv. shlukování. Další metodou spadající pod klasifikaci dat je například diskriminační analýza.

Holčík a Komenda (2015) klasifikační metody dělí do skupin dle různých hledisek. Například dle reprezentace vstupních dat na klasifikátory příznakové, strukturální a kombinované. V případě příznakových jsou vstupní data vyjádřena v podobě vektoru hodnot jednotlivých proměnných. Vektor hodnot může být buď zpracován paralelně jako celek. V takovém případě se jedná o klasifikaci paralelní, na které je založena např. metoda Bayesova klasifikátoru. Nebo může být vektor hodnot zpracován postupně neboli sekvenčně. Na tomto principu je založena např. metoda zvaná klasifikační strom. Strukturální klasifikátory, nazývané někdy té syntaktické, jsou založené na popisu jednotlivých elementárních částí vstupních dat a relací mezi nimi.

Jako další možné dělení klasifikačních metod uvádí Holčík a Komenda (2015) je jejich rozdělení na základě jednoznačností zařazení do skupin na deterministické a pravděpodobnostní. V případě deterministických je každý objekt jednoznačně zařazen do skupiny a nemůže být současně ve více skupinách. U pravděpodobnostních, též nazývaných někdy fuzzy, je získaným výsledkem pravděpodobnost zařazení objektu do jednotlivých skupin. Pojem deterministické klasifikace může mít však kromě uvedeného významu ještě další význam. Existuje i dělení klasifikace na deterministickou a nedeterministickou v tom smyslu, že deterministická metoda zpracuje data vždy se stejným výsledkem (např. Bayesův klasifikátor) a naopak nedeterministická, jež může při opakovaném zpracování určitých dat dávat různé výsledky (např. neuronové sítě).

Dle způsobů učení lze rozdělit klasifikační metody na tzv. učení s učitelem a tzv. učení bez učitele. V případě učení s učitelem jsou k dispozici trénovací data. U těchto trénovacích dat je buďto předpokládáno úplné správné rozdělení do jednotlivých skupin a v takovém případě mluvíme o učení s dokonalým učitelem. Nebo o učení s nedokonalým učitelem, kdy připouštíme, že v trénovací datech se můžou vyskytovat nesprávně rozdělené objekty do skupin. V případě učení bez učitele nejsou k dispozici žádná trénovací data a často ani nejsou známé jednotlivé skupiny, do kterých by mohla data být rozřazena. Typickým příkladem klasifikační metody založené na učení bez učitele je metoda shlukování.

3.2.4.1 Naïve Bayes

Metoda naivní Bayes či též naivní Bayesův klasifikátor (angl. Naïve Bayes) vychází z Bayesovy věty o podmíněných pravděpodobnostech. Tento klasifikační algoritmus, založený na statistické teorii o podmíněné pravděpodobnosti, je pro svou jednoduchost a kvalitní výsledky často využíván v prediktivní analýze.

Klouda a Vašata (2020) uvádí, že Bayesův klasifikátor je založen na předpokladu, že pro fixní hodnotu cílové proměnné jsou jednotlivé prediktory nezávislé. Tento předpoklad je velmi zjednodušený a nepřesný. Proto je metoda nazývána "naivní". Přesto však výsledky naivního Bayesova klasifikátoru jsou často až překvapivě dobré.

Teorie podmíněné pravděpodobnosti říká, že pravděpodobnost, že nastane jev x za podmínky, že nastal jev y je: P(x|y) = P(x,y) / P(y), kde P(x|y) je pravděpodobnost, že oba jevy nastávají současně, tedy pravděpodobnost sdružená, pro kterou platí: P(x,y) = P(x) * P(y), přičemž jevy x, y jsou nezávislé.

Bayesovu větu získáme odvozením z formulace sdružené pravděpodobnosti pro pravděpodobnost P(y|x) a zní P(y|x) = P(x|y) * P(y) / P(x).

Naivní Bayesův klasifikátor vyjádřený jako $P(y|x_{1,...,}x_i) = P(x_{1,...,}x_i) * P(y) / P(x_{1,...,}x_i)$, kde $x_{1,...,}x_i$ jsou hodnoty jednotlivých prediktorů a y je cílová proměnná, vychází z předpokladu, že jevy $x_{1,...,}x_i$ jsou nezávislé a hledá jev $P(y|x_{1,...,}x_i)$ s nejvyšší pravděpodobností. Algoritmus počítá příslušné pravděpodobnosti na základě četností výskytů hodnot v trénovacích datech.

Jednoduchost metody Naïve Bayes má však i své nevýhody. V případě, kdy by se v trénovacích datech určitá hodnota prediktoru nevyskytovala, pak by podmíněná pravděpodobnost, na které je metoda postavena, byla rovna nule. Poté by v důsledku násobení s ostatními pravděpodobnostmi byly i tyto stanoveny jako nulové a výsledky by byly tímto zkreslené. Z důvodu zamezení tomuto jevu je využíváno různých korekcí. Například korekce Laplaceovy, která spočívá v přičtení nenulových hodnot k hodnotám z nichž jsou pravděpodobnosti vypočítávány, čímž se zamezí, aby výsledná pravděpodonost byla nulová (Kohavi et al. 1997).

3.2.4.2 Logistická regrese

V případě analýzy dat, v nichž je cílová proměnná binárního typu, je využíván též model logistické regrese. Binární proměnná je proměnná nabývající právě dvou hodnot, které zpravidla vypovídají o tom, zda sledovaný jev nastal či nenastal a jsou typicky reprezentovány hodnotami 0 a 1.

Budínová (2017) logistickou regresi popisuje jako metodu založenou na odhadování pravděpodobnosti nastoupení určitého jevu. Metoda k tomuto využívá známých

vysvětlujících proměnných (prediktorů), které mohou být jak typu kvalitativního, tak kvantitativního. Za účelem zajištění predikovaných pravděpodobnostní v požadovaném intervalu (0, 1) je zaváděn pojem tzv. šance, která vyjadřuje, kolikrát je při daném prediktoru vyšší pravděpodobnost nastání jevu než jeho nenastání. Šance nabývá hodnot z intervalu (0, +∞) a následně je pomocí přirozeného logaritmu transformována na interval (-∞, +∞). Tato transformace se nazývá logitová, zkráceně logit. Logit je již možné modelovat pomocí lineárního regresního modelu.

Budínová (2017) upozorňuje na skutečnost, že pravděpodobnost nastání jevu, šance a logit jsou tři různé způsoby vyjádření téže věci. Pravděpodobnost, šance a logit jsou na sebe navzájem převoditelné.

Řeháková (2000) a rovněž Meloun a Militký (2012) logistickou regresi popisují jako metodu spočívající v logitové trasformaci založené na tzv. poměru šancí, přičemž šanci definují jako poměr pravděpodobnosti výskytu sledovaného jevu a doplňku této pravděpodobnosti, tj. pravděpodobnosti, že sledovaný jev nenastane. Proměnná logit je přirozeným logaritmem šance a její hodnoty se pohybují v intervalu od mínus nekonečna do plus nekonečna.

Princip logitové transformace spočívá v poměru šancí. Šance je podíl pravděpodobnosti, že jev nastane (Y=1) a pravděpodobnosti, že jev nenastane (Y ≠ 1), tedy
šance (Y = 1) = P (Y = 1) / [1 – P (Y = 1)].
Přirozeným logaritmem šance získáme proměnnou, jež se nazývá logit. Hodnoty proměnné logit se mohou pohybovat od mínus nekonečna do plus nekonečna:
logit (Y) = ln {P (Y = 1) / [1 – P (Y = 1)]}.
Regresní rovnice má tvar logit (Y) = $\beta_0 + \beta_1 X_1 + \ldots + \beta_i X_i$.
Logit lze převést opět zpět na šanci využitím exponenciální funkce:
šance (Y = 1) = exp [logit (Y)].
Šanci je možné převést zpět na pravděpodobnost:
P (Y = 1) = šance (Y = 1) / [1 + šance (Y = 1)]
(Řeháková, 2000, Meloun a Militký, 2012).

Budínová (2017) k parametru β_0, představujícímu absolutní člen, uvádí, že udává velikost logitu pro nulové hodnoty všech vysvětlujících proměnných. V případě, kdy je parametr β_0 roven 0, je logaritmus šance 0 a tedy šance = 1, což znamená, že pravděpodobnost nastání jevu je 0,5. Pro $\beta_0 > 0$ je šance nastání jevu větší než 0,5 a pro $\beta_0 < 0$ je šance nastání jevu menší než 0,5. K interpretaci parametrů vysvětlujících proměnných Budínová (2017) uvádí, že jednotková změna parametru β_i znamená e^{bi} násobnou změnu šance nastání jevu, za předpokladu, že ostatní proměnné zůstanou nezměněné.

K výběru proměnných vhodných pro modelování Budínová (2017) uvádí, že do modelování jsou zahrnovány takové proměnné, u nichž výsledná p-hodnota testu významnosti není vyšší než hladina významnosti 0,05. Rovněž Abbott (2014) p-hodnotu popisuje jako nástroj k vyhodnocení, zda proměnnou do modelování zahrnout či nikoli. Uvádí, že jako kritérium k zařazení určité proměnné do modelu slouží stanovení hladiny významnosti, tj. konkrétní p-hodnoty, přičemž jako typické uvádí p-hodnotu menší než 0,05 pro zařazení proměnné a p-hodnotu větší než 0,1 pro vyřazení proměnné.

Abbott (2014) jako podmínku aplikace logistické regrese uvádí, že všechny vysvětlující proměnné musí být číselného typu. Proto musí být kvalitativní proměnné transformovány na numerické pomocí dummy proměnných. Zpravidla jsou pro tento účel využívány dummy proměnné tvořené hodnotami 0 a 1. Pro každou kategorii kvalitativní proměnné je vytvořena jedna dummy proměnná.

Parametry modelu logistické regrese se odhadují. Existuje více metod, pomocí nichž je možné odhadnout parametry logistické regrese, například metoda maximální věrohodnosti. Tato metoda, jak uvádí Dokoupil (2012), spočívá v tom, že odhady parametrů náhodné veličiny se zvolí takové, aby hustota pravděpodobnostní funkce byla při těchto odhadech maximální.

3.2.4.3 Rozhodovací strom

Rozhodovací strom je často volenou metodou při řešení úloh klasifikačních i regresních. Dle typu úlohy resp. cílové proměnné lze rozlišit rozhodovací strom regresní a rozhodovací strom klasifikační.

Rozhodovací stromy je tvořeny rozhodovacími pravidly, která jsou hierarchicky uspořádána. Rozhodovací strom se skládá z kořene, který je pouze jeden, a postupně se dělí do dalších uzlů. V případě, kdy se uzel již dále nedělí, nazývá se uzlem terminálním nebo také listem. Dle způsobu dělení uzlů lze rozhodovací stromy dělit na binární a nebinární. V případě binárních se uzly dělí na dvě větve. V případě uzlů nebinárních se uzly dělí na více než dvě větve. (Komprdová, 2012)

Abbott (2014) i Dean (2014) uvádí metody založené na stromových hierarchických strukturách jako jedny z nejpopulárnějších při analýze dat. Do této kategorie spadají i tzv. rozhodovací stromy (angl. decision trees). Jako důvody široké oblíbenosti rozhodovacích stromů Dean (2014) uvádí jejich jednoduchou aplikaci i interpretaci, při současné vysoké úspěšnosti predikce.

Abbott (2014) jako hlavní výhody rozhodovacích stromů uvádí jejich snadnou pochopitelnost, jelikož je možné je číst jako sadu pravidel generujících predikovanou hodnotu. Dále zmiňuje snadné sestavení rozhodovacího stromu a též skutečnost, že s větším počtem zpracovávaných záznamů roste úspěšnost predikce rozhodovacích stromů. Jsou tudíž vhodné pro analýzu big data.

Larose (2015) též uvádí jako nejatraktivnější vlastnost rozhodovacího stromu jeho snadnou interpretaci a zdůrazňuje snadné vytváření rozhodovacích pravidel z rozhodovacího stromu. Tato pravidla mohou být jednoduše odvozena z rozhodovacího stromu postupným průchodem od kořene stromu do vybraného terminálního uzlu neboli listu.

Jako velkou výhodu rozhodovacích stromů Abbott (2014) uvádí rovněž jejich schopnost pracovat s numerickými i kategorickými proměnnými zároveň. Na rozdíl od jiných metod,

které často vyžadují, aby všechny prediktory byly buď pouze v podobě kategorické anebo pouze v podobě numerické, jako je tomu například u lineární regrese či neuronových sítí.

Jako další důvod, proč při volbě metod pamatovat na rozhodovací strom, Abbott (2014) jmenuje schopnost rozhodovacího stromu vybírat podstatné proměnné. Lze jej proto využít během fáze přípravy dat v případě, kdy je k dispozici mnoho proměnných a je potřeba zjistit, které z nich jsou pro predikci významné.

Abbott (2014) i Komprdová (2012) uvádí jako významnou výhodu rozhodovacích stromů skutečnost, že nemají požadavky na splnění určitých předpokladů, jako je tomu u některých jiných metod. Podmínkou aplikace tedy není ani požadavek na splnění předpokladu o normálním rozložení, ani předpoklad o konstantním rozptylu či nezávislosti prediktorů. Rozhodovací strom může být tudíž využit u mnoha datových souborů bez nutnosti transformace proměnných. Proměnné nemusí splňovat určité rozdělení a mohou obsahovat odlehlé hodnoty.

Abbott (2014) v neposlední řadě zmiňuje jako výhodu rozhodovacích stromů schopnost poradit si s chybějícími hodnotami v datech automaticky, aniž by bylo třeba je před sestavením modelu ošetřovat.

Výše uvedené však neznamená, že by na rozhodovací stromy nebyly kladeny vůbec žádné požadavky. Pro úplnost je třeba uvést, že vzhledem k tomu, že rozhodovací stromy jsou metodou tzv. supervizovanou či tzv. metodou s učitelem, vyžadují pro svou funkčnost data, na kterých se budou moci učit. Larose (2015) uvádí, že v případě rozhodovacího stromu musí být splněn požadavek na dostupná trénovací data. Tato data by měla poskytovat dostatek různorodých záznamů. Rozhodovací strom se učí na příkladech, tudíž, nebyla-li by určitá kategorie v datech zastoupena, docházelo by k horší úspěšnosti predikce.

Jako další požadavek na rozhodovací stromy Larose (2015) uvádí diskrétní, nikoli spojitou, cílovou proměnnou, což odůvodňuje tím, že cílová proměnná musí nabývat hodnot, které jsou od sebe navzájem dobře odlišitelné, aby šlo jednoznačně rozhodnout, zda náleží do určité kategorie či nikoliv.

Komprdová (2012) rozhodovací stromy dělí dle typu cílové proměnné na stromy klasifikační a regresní. U klasifikačního stromu je kategoriální cílová proměnná rozřazována do jednotlivých kategorií. V případě regresního stromu je spojitá závisle proměnná rozdělena do uzlů na základě hodnot prediktorů. V případě, kdy by se vyskytovaly prediktory kategoriální, jsou hodnoty rozřazovány dle kategorií prediktorů. Při větvení rozhodovacího stromu je využit vždy jen jeden prediktor. Tento stejný prediktor však může být využit i v jiném větvení. Každá hodnota prediktoru náleží pouze do jednoho terminálního uzlu neboli listu. Na základě příslušného terminálního listu je dané hodnotě určena buďto kategorie (v případě klasifikačního stromu) nebo průměr hodnot daného uzlu (v případě regresního stromu).

Dle Komprdové (2012) algoritmů pro tvorbu rozhodovacích stromů existuje vícero. Jedním z nejvyužívanějších patří algoritmus CART (Classification and Regression Trees Algorithm), který je určen pro tvorbu rozhodovacích stromů binárních. Pro nebinární rozhodovací stromy jsou určeny například rozhodovací stromy typu CHAID, které jsou určené pro proměnné kategoriální a ordinální. Pro regresní problémy je určen například rozhodovací strom PRIM či MARS. V případě algoritmu PRIM je výsledkem sada rozhodovacích pravidel bez stromové struktury. V případě metody MARS je výsledkem regresní rovnice.

Larose (2015) rozhodovací strom typu CART popisuje jako striktně binární, obsahující přesně dvě větve pro každý uzel. Naproti tomu o rozhodovacím stromu typu C4.5 uvádí, že není striktně binární, tj. uzel může mít více než dvě větve. A jako pozoruhodnou vlastnost uvádí u algoritmu C4.5, že v případě kategorických proměnných tento typ rozhodovacího stromu v základní podobě vytváří oddělené větvě pro každou hodnotu kategorické proměnné, což v některých případech může být výhodné, ale v jiných naopak nevýhodné.

Klouda a Vašata (2020) ke konstrukci rozhodovacích stromů uvádí, že se během ní používá tzv. hladový algoritmus, označovaný jako ID3 resp. jeho modifikace C4.5 a C5. Algoritmus spočívá v postupném rozdělování dat na dvě části na základě vysvětlujících proměnných, přičemž algoritmus vybírá z proměnných takovou, pomocí které je docíleno

nejvyššího informačního zisku (angl. information gain). Informační zisk odráží míru uspořádanosti rozdělené množiny dat, přičemž tato uspořádanost by byla maximální, v případě kdy by se podařilo data rozdělit způsobem, že by v jedné části byly pouze 0 a v druhé části pouze 1. Naopak uspořádanost by byla minimální v případě, kdy by byly po rozdělení dat na dvě části v obou vzniklých množinách hodnoty 0 i 1 zastoupeny ve stejné míře, tudíž by byl informační zisk minimální. Na vzniklé množiny je opakovaně aplikován stejný algoritmus, na základě kterého jsou dále děleny na další množiny, a to do té doby, dokud nenastane tzv. zastavovací kritérium, jímž může být například dosažení přednastavené maximální hloubky stromu.

K validaci modelu rozhodovacího stromu Klouda a Vašata (2020) uvádí, že se využívá ověření na testovacích datech. Před samotnou konstrukcí modelu jsou dostupná data rozdělena na dvě části: na část trénovací, na které se bude model učit a pomocí níž bude tudíž vytvořen, a na část testovací, obsahující data, pro model neznámá, na níž bude ověřena predikční schopnost modelu. Pomocí tohoto postupu je možné se co nejvíce přiblížit situaci, kdy bude model aplikován na zcela nová data. Trénovací chyba vyjadřuje poté chybovost modelu při vyhodnocení na trénovacích datech. Testovací chyba vyjadřuje chybovost při použití na testovacích datech. Obecně platí, že čím hlubší strom, tím menší trénovací chyba. Ovšem podstatně důležitější je chyba testovací. V případě, kdy bychom měli nízkou trénovací chybu a naopak vyšší testovací, je pravděpodobné, že došlo k tzv. přeučení (angl. overfitting), tj. situaci, kdy byl vytvořen strom natolik složitý, že se podařilo dosáhnout velmi nízké trénovací chyby, ovšem nedošlo k nalezení skutečně fungujícího modelu, jež by dobře predikoval data. Obecně platí, že testovací chyba s rostoucí složitostí stromu nejprve klesá, avšak v určitém okamžiku začne opět narůstat. Přičemž právě v bodě, ve kterém je testovací chyba nejmenší, se nachází nejvhodnější model.

3.2.4.4 Náhodný les

K metodě nazývané náhodný les Klouda a Vašata (2020) uvádí, že spadá do kategorie tzv. ensemble metod. Základem těchto metod je vytvoření vícero modelů stejného typu a následné zkombinování jejich výsledků do konečné predikce. Silnou stránkou těchto metod je, že v případě, dávají-li jednotlivé samostatné modely slabší výsledky, jejich zkombinováním se zpravidla výrazně zvýší predikční schopnost. Příkladem aplikace ensemble metod je například tzv. náhodný les (angl. random forest).

Dle Kloudy a Vašaty (2020) je u ensemble metod důležité, aby se jednotlivé modely neopakovaly neboli aby byly dostatečně odlišné. Toho se dosahuje aplikací různých metod. V případě náhodných lesů je touto metodou metoda zvaná bootstrap. Náhodné lesy jsou vytvářeny pomocí metody zvané bagging, neboli bootstrap aggregating či lze také říci pomocí výběru s opakováním. Z trénovacích dat je pomocí metody výběru s opakováním vytvořen určitý počet stejně velikých dat. Na základě jednotlivých vytvořených datových souborů, jsou vytvořeny příslušné modely rozhodovacích stromů. Tyto stromy tvoří dohromady tzv. náhodný les, přičemž výsledek predikce tohoto náhodného lesa je dán většinovým hlasováním (výsledky) jednotlivých stromů, z nichž les sestává. V případě, kdy predikce většiny stromů je například hodnota 1, je i predikce náhodného lesa hodnota 1.

Náhodný les je oproti samostatnému rozhodovacímu stromu znatelně odolnější vůči přeučení. Nevýhodou však je, že bývají většinou složité a proto i hůře interpretovatelné.

V závislosti na typu cílové proměnné se rozlišují rozhodovací lesy klasifikační a regresní.

Klasifikační les je model sestávající z určitého počtu klasifikačních stromů. Výsledná predikovaná hodnota je vytvořena na základě hlasování či jako průměr pravděpodobnostního zastoupení hodnoty v konečném uzlu stromu neboli listu.

Regresní les je model sestávající z několika regresních stromů. Výsledná predikovaná hodnota je vytvořena na základě váženého průměru regresních funkcí regresních stromů.

3.2.5 Hodnocení kvality modelu

Za účelem hodnocení kvality klasifikačních modelů se využívá aplikace vytvořeného modelu, vzniklého z učení na trénovacích datech, na data testovací, tj. data pro model dosud neznámá. Jelikož často není k dispozici další nezávislý datový soubor, využívá se rozdělení dostupných dat na data trénovací a data testovací. Ayyadevara (2018) jako typický poměr rozdělení dat na trénovací a testovací uvádí 60 % dat pro data trénovací a zbytek pro validaci a testování.

Holčík a Komenda (2015) popisují různé způsoby rozdělení dat na data trénovací a testovací. Patří k ním resubstituce (angl. resubstitution), náhodný výběr s opakováním (angl. bootstrap), predikční testování externí validací (angl. hold-out), k-násobná křížová validace (angl. k-fold cross-validation).

V případě resubstituce jsou stejná data využita jako trénovací i jako testovací. Tento postup je však nevhodný, jelikož vzniká vysoké riziko tzv. přeučení modelu. Přeučený model má velmi dobré výsledky na trénovacích datech, ale naopak velmi špatné výsledky při aplikaci na data zcela nová, nezávislá. Tato metoda se z tohoto důvodu již prakticky nepoužívá.

Náhodný výběr s opakováním spočívá v n-krát opakovaném náhodném výběru z dat, která májí n záznamů. Vybraná data se použijí jako data trénovací a nevybraná jako testovací. Nevýhodou však je, že se záznamy v trénovacích datech opakují. Při analýze big data se tento způsob zpravidla nevyužívá, jelikož v jejich případě není problém s velikostí dat.

Při predikčním testování externí validací je zpravidla větší část dat použita jako trénovací a menší část dat jako testovací. Výhodou tohoto způsobu je nezávislost vzniklých částí dat. Nevýhodou může být zmenšení datového souboru v případě menšího počtu dostupných dat. V případě tohoto způsobu však existuje riziko, že rozdělení dat bude mít za následek neodpovídající rozložení v datech, kdy by se například mohly odlehlé hodnoty, mnohdy velmi významné pro analýzu, vyskytovat pouze v datech trénovacích nebo pouze v datech testovacích. Tento způsob je možné využít při práci s big data, jelikož s rostoucím počtem dat se snižuje riziko neodpovídajícího rozdělení dat a zvyšuje se přesnost modelu.

K-násobná křížová validace data rozdělí na *k* částí. Jedna část je použita na testování a zbylé části na trénování. Tento proces se opakuje do okamžiku, kdy každá část je použita na testování právě jednou. Výhodou tohoto způsobu je přesná validace klasifikačního modelu. Pro velké datové soubory se zpravidla využívá rozdělení na 10 částí, tedy 10 násobná křížová validace (Holčík a Komenda, 2015).

Matice záměn

Při hodnocení kvality binárního klasifikačního modelu je využíván nástroj zvaný matice záměn (angl. confusion matrix), pomocí kterého je porovnáván výsledek predikce se skutečnými daty. Je hodnoceno, jak dobře byly jednotlivé záznamy vyhodnoceny z hlediska jejich zařazení do kategorií cílové proměnné.

Obrázek 3: Matice záměn

Confusion Matrix		Predicted Class		
		0 (predicted value is negative)	**1** (predicted value is positive)	Total Actual (down)
Actual Class	**0** (actual value is negative)	t_n (true negative)	f_p (false positive, false alarm)	Total actual negatives tn + fp
	1 (actual value is positive)	f_n (false negative, false dismissal)	t_p (true positive)	Total actual positives tp + fn
	Total Predicted (across)	Total negative predictions tn + fn	Total positive predictions tp + fp	Total Examples tp + tn + fp + fn

Zdroj: Abbott (2014)

Konkrétně matice záměn obsahuje četnosti správně klasifikovaných záznamů a nesprávně klasifikovaných záznamů:

- počet skutečně pozitivních (TP – true positive)
- počet falešně negativních (FN – false negative)
- počet skutečně negativních (TN – true negative)
- počet falešně pozitivních (FP – false positive).

Z matice záměn je možné získat výpočtem různá hodnocení kvality modelu. Jedná se například o celkovou správnost či přesnost (angl. accuracy), chybu (angl. error), senzitivitu (angl. sensitivity), specificitu (angl. specificity).

Celková přesnost je podíl správně klasifikovaných ze sumy všech záznamů. Chyba představuje podíl chybně klasifikovaných ze sumy všech záznamů.

Precisnost udává podíl skutečně positivních. Senzitivita či úplnost vyjadřuje podíl hodnot skutečně positivních, které se danému modelu podařilo nalézt. Specificita představuje podíl skutečně negativních, které se podařilo odhalit.

ROC křivka

Označení ROC je odvozeno od anglického názvu Receiver Operating Characteristic Curve. ROC křivka (též AUC z angl. Area Under Curve) poskytuje grafické znázornění vztahu mezi senzitivitou a specificitou. ROC křivka vychází z matice záměn. Je využívána k hodnocení kvality a k optimalizaci modelů. Rovněž je využívána k porovnání více modelů.

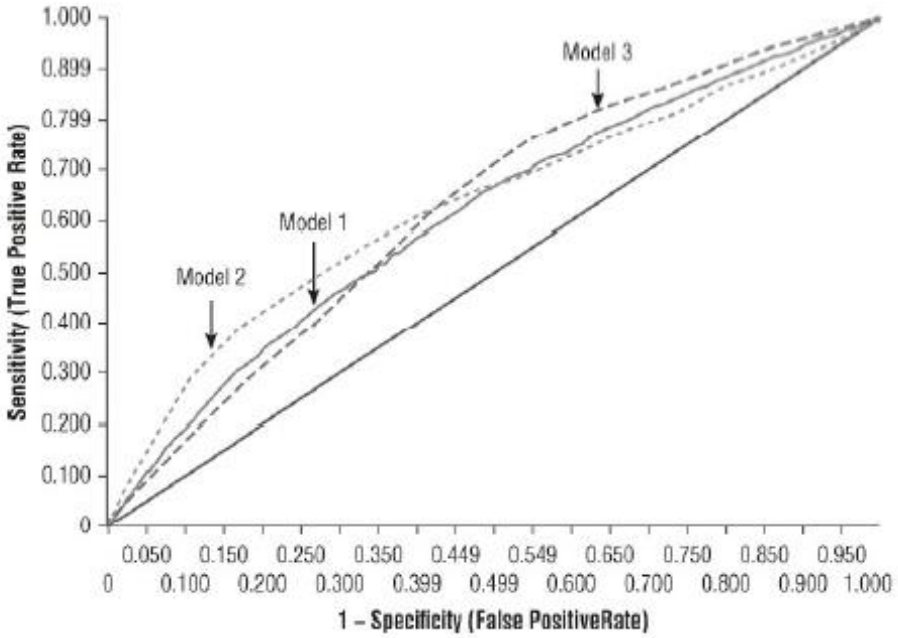
Osa y grafu představuje podíl skutečně pozitivních (TP) záznamů a osa x představuje podíl falešně pozitivních záznamů. Neboli osa y představuje pravděpodobnost, že jako správný bude vyhodnocen pozitivní záznam, a osa x pravděpodobnost, že jako správný bude vyhodnocen negativní záznam.

Body na ROC křivce představují konkrétní hodnotu tzv. dělícího kritéria (angl. treshold). Dělící kritérium představuje práh, na základě kterého je algoritmem rozhodnuto, zda přiřadí daný záznam do skupiny pozitivních či skupiny negativních. Zpravidla je používána hodnota prahu 0,5. (Abbott, 2014)

Mandrekar (2010) hodnotí ROC křivku resp. hodnotu AUC jako efektivní způsob vyhodnocení celkové přesnosti modelu. Uvádí, že může nabývat hodnoty od 0 do 1, přičemž hodnota 0 znamená úplnou nepřesnost a hodnota 1 úplnou přesnost.

Mandrekar (2010) k interpretaci hodnoty ROC resp. AUC dále uvádí, že výsledná hodnota testu 0,5 indikuje náhodný model. Hodnota v intervalu od 0,7 do 0,8 již značí přijatelnou predikci. Hodnota pohybující se v intervalu 0,8 až 0,9 znamená výbornou predikci. V případě, kdy se hodnota nachází v intervalu 0,9 až 1,0, interpretuje ji jako predikci vynikající.

Graf 1: ROC křivka – příklad srovnání více modelů



Zdroj: Abbott (2014)

3.3 Řízení lidských zdrojů

Armstrong (2007) definuje řízení lidských zdrojů jako strategický a logicky promyšlený přístup k řízení lidí, kteří jednotlivě i kolektivně přispívají k dosažení cílů organizace. Řízení lidských zdrojů dále popisuje jako propojený systém, do nějž patří jednak filozofie lidských zdrojů popisující základní principy využívané v řízení lidí, dále též strategie lidských zdrojů určující cíl řízení lidských zdrojů a rovněž i zásady a pravidla definující způsob uplatňování principů a strategií v jednotlivých oblastech řízení lidských zdrojů.

Armstrong (2007) jako obecný cíl řízení lidských zdrojů uvádí zajištění, aby organizace prostřednictvím lidí úspěšně dosahovala svých cílů. Konkrétněji uvádí, že řízení lidských zdrojů souvisí s dosahováním cílů v především v oblastech, mezi které spadá zlepšení efektivnosti organizace prostřednictvím řízení znalostí a talentů, dále řízení lidského kapitálu, představovaného lidmi pracujícími v organizaci. Lidský kapitál popisuje Armstrong (2007) jako nejcennější kapitál organizace, do kterého je třeba investovat.

Armstrong (2007) jako hlavní cíl řízení lidských zdrojů uvádí zajištění, aby organizace nalezla a zároveň si byla schopná udržet nejen kvalifikovanou, ale rovněž i motivovanou pracovní sílu. To znamená též, aby organizace dokázala předvídat potřeby pracovníků, usilovat o jejich uspokojování a též investovat do zvyšování schopností lidí k jejich rozvoji. Rovněž to znamená nastavit postupy získávání a výběru pracovní síly, vzdělávacích aktivit a systému odměňování.

Armstrong (2007) zdůrazňuje, že postupy a procesy uplatňované v rámci řízení lidských zdrojů jsou ovlivňovány nejen podmínkami vnitřními, ale také výrazně závisí na faktorech vnějších. Jako hlavní externí faktory uvádí technologie a konkurenční tlaky. Technologie má zásadní vliv na vnitřní prostředí organizace, na to jakým způsobem je práce prováděna a organizována. V případě, kdy dojde ke změně technologie, změní se související procesy i požadavky na pracovní sílu. Může to vést k rozšíření počtu pracovních sil v organizaci a jejich víceoborovosti, ale může být také příčinou dekvalifikace a snížení počtu pracovních míst. Konkurenční tlaky též vedou k zavádění nových technologií.

3.3.1 Plánování lidských zdrojů

Armstrong (2007) oblast zabezpečování lidských zdrojů uvádí jako klíčovou pro získání a udržení pracovních sil. Tato oblast sestává z plánování lidských zdrojů, řízení talentů a získávání lidských zdrojů. Plánování lidských zdrojů určuje lidské zdroje potřebné k dosažení cílů organizace. Rozlišuje tvrdé a měkké plánování lidských zdrojů. Tvrdé spočívá v kvantitativní analýze za účelem určení kolik pracovních sil je zapotřebí. Měkké spočívá v zajišťování, aby pracovní síla sestávala z lidí motivovaných pro danou práci. Měkké metody využívají zpravidla dotazníkových šetření mezi pracovníky.

Dle Armstronga (2007) plánování lidských zdrojů sestává ze tří fází, kterými jsou předvídání poptávky po pracovní síle, předvídání nabídky pracovní síly a sestavování plánů majících za účel sladit nabídku a poptávku po pracovní síle.

K oblasti plánování lidských zdrojů Armstrong (2007) zmiňuje, že mezi teorií a praxí existuje propast, jelikož je poměrně obtížné předvídat budoucnost. Též zmiňuje určitou přetrvávající nedůvěru manažerů v teorii plánování.

Předvídání budoucnosti se však stává zásluhou metod, mezi které patří prediktivní modelování, snazší a rozšiřováním aplikace těchto metod v praxi se postupně snižuje i nedůvěra v plánování.

Samotný proces plánování lidských zdrojů sestává z podprocesů, mezi něž patří v prvé řadě podnikové strategické plány, v nichž jsou stanoveny požadavky na pracovní síly. Dále strategie zabezpečování lidských zdrojů, která zahrnuje plánování v rámci zajištění konkurenční výhody, tj. získání pracovních sil lepších než kterými disponuje konkurence. Dalším procesem plánování lidských zdrojů jsou prognózy poptávky a nabídky, prezentované jako odhady potřebného počtu a kvalifikace pracovních sil a též odhady počtu dostupné pracovní síly uvnitř i vně organizace. Dalšími procesy plánování lidských zdrojů jsou analýza pohybu neboli fluktuace zaměstnanců, sestávající z analýzy faktorů ovlivňujících fluktuaci. Součástí plánování je i analýza pracovního prostředí a analýza produktivity.

3.3.2 Odchody zaměstnanců

Armstrong (2007) uvádí, že zkoumáním počtů zaměstnanců odcházejících z organizace je možné získat údaje využitelné při výpočtech budoucích ztrát zaměstnanců. Zdůrazňuje též, že analýzou odchodů zaměstnanců a příčin těchto odchodů získá organizace cenné informace týkající se možné potřeby učinění kroků za účelem snížení míry odchodů zaměstnanců. Může přinést poznatky, na které by bylo vhodné se blíže zaměřit a zkoumat je a určit časté faktory ovlivňující odchod zaměstnanců a odhalit možné řešení k odstranění či zmírnění příčin.

Armstrong (2007) uvádí, že vysoká míra odchodů zaměstnanců z organizace může vést k její destabilizaci. Též zvyšuje náklady organizace, mezi kterými uvádí jednak náklady mzdové a administrativní, přímo spojené s odchodem zaměstnance. Dále náklady spojené se získáváním náhrady za odcházejícího zaměstnance, které zahrnují inzerci, pohovory a testování. Jako další uvádí náklady času, který personalisté a manažeři stráví výběrem nového zaměstnance a následně času, jenž stráví jeho zapracováním a zaučováním, náklady na vstupní kurzy, školení a materiál. Nákladem jsou i ztráty způsobené dočasnou neobsazeností daného pracovního místa.

Faktory ovlivňující odchody zaměstnanců

Dle Armstronga (2007) analýza faktorů ovlivňujících odchody zaměstnanců poskytuje cenné informace, jež je následně možné využít k zvýšení stability zaměstnanců.

Příčiny odchodů rozděluje do následujících skupin:
- nabídka vyšší mzdy v jiné organizaci,
- vyhlídka na lepší vývoj kariéry v jiné organizaci,
- vyšší jistota zaměstnání v jiné organizaci,
- příležitosti více rozvíjet dovednosti a kvalifikaci v jiné organizaci,
- lepší pracovní podmínky v jiné organizaci,
- špatné vztahy s nadřízeným,
- špatné vztahy s kolegy,

- zastrašování či obtěžování,
- osobní důvody, mezi které patří například stěhování apod.

S odcházejícím pracovníkem by dle Armstronga (2007) měl být veden rozhovor za účelem zjištění důvodu odchodu.

Armstrong (2007) v návaznosti na analýzu odchodů zaměstnanců uvádí jako následné možné preventivní kroky, mezi kterými uvádí především hodnocení systému odměn, který by měl být konkurenceschopný. Dále hodnocení způsobu vytváření pracovních míst, přičemž pracovní místo by mělo nabízet rozmanitost práce, významnost úkolů, zpětnou vazbu, rozšiřování dovedností a možnost růstu. Dále hodnocení angažovanosti v práci v souvislosti s organizováním projektů, s nimiž se zaměstnanci mohou lépe identifikovat. Rovněž navrhuje hodnocení vytváření sociálních vazeb v organizaci, hodnocení rovnováhy mezi pracovním a soukromým životem zaměstnanců, hodnocení existence nepříznivých pracovních podmínek. Též doporučuje zaměřit se informování a vzdělávání vedoucích zaměstnanců za účelem správného vedení svých zaměstnanců.

Analýzou faktorů mající vliv na odchod zaměstnanců za zabývá studie autorů Srivastava a Tiwari (2020). Jako faktory nejvíce ovlivňující míru odchodů zaměstnanců uvádí časté pracovní cesty, velkou vzdálenost z domova do práce, dále skutečnost, že zaměstnanec má za sebou vícero zaměstnání v různých společnostech, dále nutnost přesčasů k dokončení práce. Naopak jako faktory nejvíce snižující míru odchodů zaměstnanců uvádí spokojenost s pracovním prostředím, spokojenost s náplní práce a angažovanost v zaměstnání.

4 Vlastní práce

Náplní praktické části práce je srovnání predikční schopnosti vybraných modelů. Dále též identifikace faktorů nejvíce ovlivňujících odchody zaměstnanců.

Praktická část práce je založena na zkoumání dat týkajících se odchodů zaměstnanců ze zaměstnání. Odchody se přitom rozumí situace, kdy se zaměstnanec rozhodne opustit organizaci na základě svého rozhodnutí. Nikoli na základě nároku na odchod do důchodu nebo na základě výpovědi dané zaměstnavatelem.

Vzhledem k dostupnosti historických dat, na základě kterých se mohou modely učit, spadá daný problém do oblasti tzv. metod učení s učitelem. Vzhledem k charakteru cílové proměnné, která je kategorická, binární, se jedná o problém klasifikační. Vybranými modely jsou klasifikační modely uvedené v teoretické části práce, konkrétně naivní Bayesův klasifikátor, logistická regrese, rozhodovací strom a náhodný les.

Analyzovaný datový soubor obsahuje 8 proměnných v 15 tisíci záznamech. Vzhledem k citlivé povaze dat jde o data plně anonymizovaná, tzn. soubor neobsahuje žádná data, jež by mohla být byť nepřímo spojena s konkrétní fyzickou či právnickou osobou či oblastí působení. Zdrojem dat je databáze webové platformy Kaggle sloužící k pořádání soutěží v oblasti datové analytiky (Kaggle, 2019).

Postup při analýze vychází z metodiky CRISP-DM, která sestává z následujících šesti fází:
1. Porozumění problematice
2. Porozumění datům
3. Příprava dat
4. Modelování
5. Vyhodnocení výsledků
6. Využití výsledků

Data jsou analyzovaná pomocí softwarového nástroje RapidMiner, verze 9.8.001, licence RapidMiner Studio Educational (Mierswa I., Klinkenberg, R., 2020).

4.1 Porozumění problematice

Odchody zaměstnanců jsou pro organizace negativním jevem, jelikož současně se ztrátou zaměstnance, ztrácejí i cenné znalosti, které zaměstnanec nabyl během působení v organizaci. Odchod zaměstnance sebou nese i nemalé náklady pro organizaci.

Z uvedených důvodů by měla být věnována zvýšená pozornost příčinám odchodů zaměstnanců. V případě, že se podaří nalézt nejčastější příčiny a odhalit tak problematické oblasti, je možné následně zavést preventivní opatření a snížit tak míru odchodů zaměstnanců z organizace.

V teoretické části byly Armstrongem (2007) jako nejčastější příčiny odchodů zaměstnanců uváděny jednak faktory týkající se odměňování zaměstnanců, dále faktory související s rozvojem pracovních dovedností a kvalifikace, a rovněž faktory týkající se pracovních podmínek a vztahů se spolupracovníky či nadřízeným. Uvedené faktory spadají do kategorie, které mohou být určitým způsobem ovlivněny ze strany zaměstnavatele. Armstrong (2007) uvádí též důvody, jakými jsou například stěhování, ty však ovlivněny zaměstnavatelem být nemohou.

Srivastava a Tiwari (2020) jako kategorie nejčastějších příčin odchodů zaměstnanců uvádí faktory týkající se počtu nutných přesčasů a častého střídání pracovních míst v pracovní historii zaměstnance. Jako faktory nejvíce snižující pravděpodobnost odchodu zaměstnance uvádí obecně jeho spokojenost a též angažovanost v zaměstnání.

K porozumění příčinám a odhalení hrozících odchodů zaměstnanců je možné využít silného nástroje v podobě prediktivního modelování, pomocí kterého je možné predikovat odchody zaměstnanců a též nalézat faktory ovlivňující odchod zaměstnanců. Následným ovlivňováním nalezených faktorů může organizace činit preventivní opatření za účelem snížení míry odchodu zaměstnanců.

Za účelem vybrání nejvhodnějšího modelu pro predikci odchodu zaměstnanců z organizace byl stanoven následující postup a cíle:

- prozkoumání dostupných dat,
- ošetření chybějících hodnot,
- transformace proměnných,
- výběr proměnných vhodných k modelování,
- rozdělení dat na část trénovací a testovací,
- sestavení predikčního modelu Naïve Bayes a vyhodnocení jeho kvality,
- sestavení predikčního modelu logistické regrese a vyhodnocení jeho kvality,
- sestavení predikčního modelu rozhodovacího stromu a vyhodnocení jeho kvality,
- sestavení predikčního modelu náhodného lesa a vyhodnocení jeho kvality,
- porovnání predikční schopnosti vytvořeních modelů srovnáním jejich hodnocení na základě matic záměn a ROC křivek,
- výběr nejvhodnějšího modelu,
- identifikace faktorů nejvíce ovlivňujících odchody zaměstnanců pomocí vybraného nejvhodnějšího modelu.

4.2 Porozumění datům

Data jsou k dispozici v souboru formátu *.csv*, což je formát určený pro tabulková data. Uložena jsou v 15000 řádcích a 11 sloupcích. Jednotlivé řádky představují konkrétní záznamy o zaměstnancích. Názvy sloupců, tj. jednotlivých proměnných, byly pro účely této práce přeloženy z anglického do českého jazyka. První sloupec obsahuje identifikátor záznamu (ID). Ostatní sloupce obsahují jednotlivé proměnné. Cílová proměnná je nazvána *odchod*. Vysvětlující proměnné neboli prediktory jsou proměnné popisující spokojenost, hodnocení, projekty, hodin měsíčně, počet let, pracovní úraz, povýšení, oddělení a mzda.

V následující tabulce (Tabulka 1) je zobrazen náhled prvních několika záznamů dat.

Tabulka 1: Ukázka dat v nezpracované podobě

Row No.	spokojenost	hodnoceni	projekty	hodin_m	pocet_let	uraz	odchod	povyseni	oddeleni	mzda
1	0.380	0.530	2	157	3	0	1	0	prodej	1
2	0.800	0.860	5	262	6	0	1	0	prodej	2
3	0.110	0.880	7	272	4	0	1	0	prodej	2
4	0.720	0.870	5	223	5	0	1	0	prodej	1
5	0.370	0.520	2	159	3	0	1	0	prodej	1
6	0.410	0.500	2	153	3	0	1	0	prodej	1
7	0.100	0.770	6	247	4	0	1	0	prodej	1
8	0.920	0.850	5	259	5	0	1	0	prodej	1

Zpracování: RapidMiner, vlastní (zdroj: Kaggle)

První sloupec je označením konkrétního řádku v datovém souboru. Jedná se tedy o hodnotu ID. Během fáze přípravy dat je zapotřebí tento sloupec z dat odfiltrovat, jelikož se jedná pouze o identifikátor jednotlivých záznamů, nikoli o proměnnou.

V druhém sloupci je obsažena proměnná *spokojenost*, která představuje výsledek posledního hodnocení spokojenosti zaměstnance. Proměnná vyjadřuje výslednou spokojenost v procentech. Nabývá hodnot od 0 do 1. Čím je hodnota proměnné vyšší, tím vyšší je spojenost zaměstnance. Tato hodnota vyjadřuje postoj daného zaměstnance.

Třetí sloupec obsahuje proměnnou *hodnocení*, která představuje výsledek hodnocení zaměstnance ze strany organizace. Proměnná je vyjádřena v procentech. Nabývá hodnot od 0 do 1. Čím je hodnota proměnné vyšší, tím je lepší hodnocení zaměstnance.

Čtvrtý sloupec představuje proměnnou *projekty*, která vypovídá o celkovém počtu projektů, do nichž byl daný zaměstnanec zapojen za dobu působení u společnosti.

Sloupec pátý obsahuje proměnnou *hodin_m* reprezentující zaměstnancův průměrný počet odpracovaných hodin v měsíci.

Proměnná *pocet_let* ve sloupci šestém, představuje zaměstnancův počet odpracovaných let u organizace.

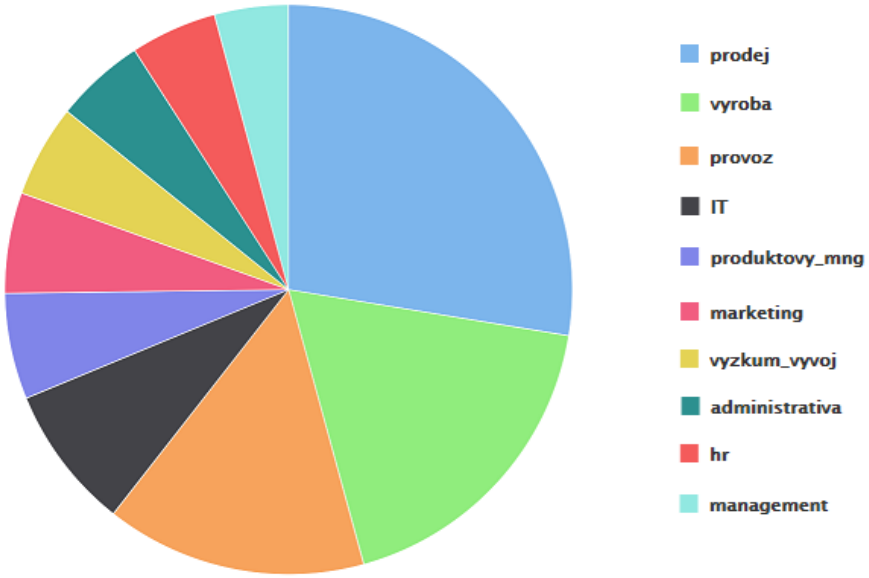
Proměnná *uraz* ve sloupci sedmém, obsahuje údaj o tom, zda zaměstnanec měl za dobu sledovaného působení u společnosti pracovní úraz či nikoliv. Jedná se o proměnnou, která může nabývat pouze dvou hodnot, tj. hodnoty 1 v případě, že se u něj pracovní úraz vyskytl, a hodnoty 0 v případě, se u zaměstnance pracovní úraz nevyskytl. Jde tudíž o proměnnou kategoriálního typu, která je binární. Během fáze přípravy dat bude ověřeno, zda tato proměnná je jako binární reprezentována.

Cílovou proměnnou je proměnná *odchod*, vyskytující se v osmém sloupci. Vypovídá o tom, zda u konkrétního zaměstnance odchod ze zaměstnání nastal či nenastal. Jedná se proměnnou, která může nabývat pouze dvou hodnot, konkrétně hodnoty 1 v případě nastání odchodu a hodnoty 0 v případě nenastání odchodu. Jde tudíž o proměnnou kategorickou binární. Během fáze přípravy dat je zapotřebí zkontrolovat, zda je proměnná *odchod* správného datového typu.

Proměnná *povyseni* v devátém sloupci, představuje informaci o tom, zda byl zaměstnanec za dobu svého působení u společnosti povýšen či nikoliv. Tato proměnná nabývá pouze dvou možných hodnot. Hodnoty 1 v případě, že povýšen byl, a hodnoty 0 v případě, že k povýšení nedošlo. Proměnná je tudíž kategorická a binární. V rámci fáze přípravy dat bude ověřeno, zda je tato proměnná binárního datového typu.

Proměnná *oddeleni* obsahuje informaci o pracovním zařazení zaměstnanců do konkrétních oddělení. Proměnná může nabývat deseti různých hodnot. Jedná se o proměnnou kategoriálního, nominálního typu, jelikož hodnoty proměnné nelze seřadit, resp. nelze je mezi sebou porovnávat. V následující fázi přípravy dat bude ověřen kategorický typ proměnné.

Graf 2: Zastoupení hodnot proměnné oddělení



Zpracování: RapidMiner, vlastní (zdroj: Kaggle)

Poslední proměnnou, kterou datový soubor obsahuje, je proměnná *mzda*, odrážející výši mzdy. Tato proměnná neobsahuje průměrné mzdy zaměstnanců a její hodnoty ani neodpovídají průměrným hodnotám v případě, že by se jednalo o již v původních datech aplikovaný binning dat. Hodnoty mzdy jsou reprezentovány číselnými hodnotami 1, 2, 3, tudíž se jedná o škálu typu nízká – střední - vysoká, přičemž ale nejsou k dispozici konkrétní hodnoty představující hranice mezi jednotlivými kategoriemi. Vzhledem k tomu, že v teoretické části práce, v kapitole o přípravě dat, bylo na základě doporučení Larose (2015) zjištěno, že není vhodným řešením kategorické proměnné transformovat do číselných proměnných, byť jsou ordinálního typu, bude v následující fázi analýzy, přípravě dat, ověřeno, zda tato proměnná je reprezentována jako kategorická.

Přehled základních popisných statistických charakteristik proměnných softwarového nástroje RapidMiner nabízí informace o typu proměnné a počtu chybějících hodnot. V případě proměnných číselných udává rovněž informace o minimální, maximální a průměrné hodnotě a směrodatné odchylce u číselných proměnných. V případě kategorických proměnných poskytuje informace o nejčastěji a nejméně často se vyskytující hodnotě.

Následující tabulka (Tabulka 2) obsahuje statistické charakteristiky nezpracovaných dat, tj. dat, která dosud neprošla žádnými úpravami a jsou v podobě, v jaké byla importována.

Tabulka 2: Statistické charakteristiky proměnných – nezpracovaná data

Name	Type	Missing	Statistics			
hodin_m	Real	0	Min 96	Max 310	Average 201.050	Deviation 49.943
hodnoceni	Real	0	Min 0.360	Max 1	Average 0.716	Deviation 0.171
mzda	Real	0	Min 1	Max 3	Average 1.595	Deviation 0.637
odchod	Real	0	Min 0	Max 1	Average 0.238	Deviation 0.426
oddeleni	Nominal	0	Least management (630)	Most prodej (4140)		Values prodej (4140), vyroba (2720), provoz (2229), IT (1227), ...[6 more]
pocet_let	Real	0	Min 2	Max 10	Average 3.498	Deviation 1.460
povyseni	Real	0	Min 0	Max 1	Average 0.021	Deviation 0.144
projekty	Real	0	Min 2	Max 7	Average 3.803	Deviation 1.233
spokojenost	Real	0	Min 0.090	Max 1	Average 0.613	Deviation 0.249
uraz	Real	0	Min 0	Max 1	Average 0.145	Deviation 0.352

Zpracování: RapidMiner, vlastní (zdroj: Kaggle)

Z uvedeného přehledu (Tabulka 2) vyplývá, že mimo proměnnou *oddeleni*, obsahující deset různých hodnot v podobě textu, všechny ostatní proměnné byly softwarovým nástrojem RapidMiner automaticky vyhodnoceny jako proměnné číselné. Je tomu tak vzhledem k obsahu proměnných, který je ve všech uvedených případech číselný.

Avšak proměnné *odchod*, *povyseni* a *uraz* jsou proměnné též kategorické, jelikož popisují, zda v daném případě příslušný jev nastal či nenastal. Jedná se tedy o kategorické proměnné binární a v další fázi, přípravě dat, je zapotřebí tyto proměnné transformovat z číselných na kategorické binární.

Rovněž proměnná *mzda* bude transformována na kategorickou, jelikož tato proměnná, jak již bylo zmíněno výše, představuje kategorickou proměnnou ordinálního typu. Tzn. její hodnoty je sice možné seřadit od nejnižší po nejvyšší a tudíž by mohla být reprezentována jako proměnná číselná. Avšak vzhledem k poznatkům nabytým v rešeršní části práce v kapitole věnované přípravě dat, kdy Larose (2015) nedoporučuje transformovat kategorické proměnné do číselných z důvodu možného zkreslení spočívajícího v možném neodpovídajícím vzájemném poměru hodnot, bude s touto proměnnou nakládáno jako s kategoriální.

Proměnné *hodin_m*, *hodnoceni*, *pocet_let*, *projekty* a *spokojenost* byly správně identifikovány jako proměnné číselné.

Rovněž proměnná *oddeleni* byla správně rozpoznána jako kategorická.

V datech se nenachází žádné chybějící hodnoty, tudíž není třeba v následující fázi přípravy dat aplikovat žádnou z možných metod jejich ošetření.

4.3 Příprava dat

Během fáze přípravy dat byla nejprve provedena transformace proměnných na základě požadavků vzniklých v předchozí fázi porozumění datům. Následně byl proveden výběr proměnných vhodných pro modelování. Vzhledem k tomu, že v předchozí fázi, porozumění datům, bylo zjištěno, že data neobsahují žádné chybějící hodnoty, nebylo tudíž v této fázi analýzy dat zapotřebí aplikace metod sloužících k ošetřování chybějících hodnot.

Transformace proměnných

Proměnné *odchod*, *povyseni* a *uraz* byly transformovány na proměnné kategorické binární, jelikož všechny nabývají pouze dvou možných hodnot: hodnoty 1 v případě nastání jevu a hodnoty 0 v případě jeho nenastání.

Následně byla též proměnná *mzda* transformována na kategoriální, vzhledem k poznatkům zjištěným v rešerši v části věnované přípravě dat, kde Larose (2015) doporučuje s kategoriálními proměnnými i v případě, kdy jsou ordinálního typu, pracovat jako s kategoriálními a netransformovat je do číselných.

Následující tabulka (Tabulka 3) obsahuje ukázku záznamů dat po fázi přípravy dat.

Tabulka 3: Ukázka dat ve zpracované podobě

hodin_m *Number*	hodnoceni *Number*	mzda *Category*	oddeleni *Category*	pocet_let *Number*	projekty *Number*	spokojenost *Number*	uraz *Category*	odchod *Category*
157	0.530	1	prodej	3	2	0.380	0	1
262	0.860	2	prodej	6	5	0.800	0	1
272	0.880	2	prodej	4	7	0.110	0	1
223	0.870	1	prodej	5	5	0.720	0	1
159	0.520	1	prodej	3	2	0.370	0	1
153	0.500	1	prodej	3	2	0.410	0	1
247	0.770	1	prodej	4	6	0.100	0	1
259	0.850	1	prodej	5	5	0.920	0	1

Zpracování: RapidMiner, vlastní (zdroj: Kaggle)

Následující tabulka (Tabulka 4) obsahuje základní statistické charakteristiky proměnných po zpracování, tj. po provedených úpravách ve fázi přípravy dat.

Tabulka 4: Statistické charakteristiky proměnných – zpracovaná data

Name	Type	Missing	Statistics			
hodin_m	Real	0	Min 96	Max 310	Average 201.050	Deviation 49.943
hodnoceni	Real	0	Min 0.360	Max 1	Average 0.716	Deviation 0.171
mzda	Nominal	0	Least 3 (1237)	Most 1 (7316)		Values 1 (7316), 2 (6446), 3 (1237)
odchod	Binominal	0	Negative 0	Positive 1		Values 0 (11428), 1 (3571)
oddeleni	Nominal	0	Least management (630)	Most prodej (4140)		Values prodej (4140), vyroba (2720), provoz (2229), IT (1227), ...[6 more]
pocet_let	Real	0	Min 2	Max 10	Average 3.498	Deviation 1.460
povyseni	Binominal	0	Negative 0	Positive 1		Values 0 (14680), 1 (319)
projekty	Real	0	Min 2	Max 7	Average 3.803	Deviation 1.233
spokojenost	Real	0	Min 0.090	Max 1	Average 0.613	Deviation 0.249
uraz	Binominal	0	Negative 0	Positive 1		Values 0 (12830), 1 (2169)

Zpracování: RapidMiner, vlastní (zdroj: Kaggle)

Výběr proměnných

Software RapidMiner poskytuje nástroj pro výběr vysvětlujících proměnných neboli prediktorů vhodných k modelování, který k vyhodnocení využívá tzv. kritérií CISMT. Kritérium C (*correlation*) je vyhodnocováno na základě korelace dané proměnné s proměnnou cílovou. Kritérium I (*ID-ness*), jak moc se daná proměnná blíží charakterem hodnot k proměnné typu identifikátor, tzn. takové, která má všechny nebo téměř všechny hodnoty unikátní. Kritérium S (*stability*) naopak uvádí, do jaké míry jsou hodnoty proměnné shodné, tj. zda se neblíží či dokonce není proměnnou unární. Kritérium M (*missing*) uvádí procento chybějících hodnot. Kritérium T (*text-ness*) zkoumá proměnné z hlediska, zda sestává z více než jednoho slova a rovněž zohledňuje délku hodnoty proměnné. Symbol ? představuje chybějící hodnotu (Manuaba et al. 2020). Hodnocení kvality prediktorů je uvedeno v následující tabulce (Tabulka 5).

Tabulka 5: Hodnocení kvality prediktorů dle kritérií CISMT

Status	Quality	Name	Correlation	ID-ness	Stability	Missing	Text-ness
●		povyseni	0.38%	0.01%	97.87%	0.00%	0.45%
●		hodnoceni	0.00%	?	2.39%	0.00%	0.00%
●		spokojenost	15.08%	?	2.39%	0.00%	0.00%
●		projekty	0.06%	0.04%	29.10%	0.00%	0.00%
●		hodin_m	0.51%	1.43%	1.02%	0.00%	0.00%
●		pocet_let	2.10%	0.05%	42.96%	0.00%	0.00%
●		uraz	2.39%	0.01%	85.54%	0.00%	0.45%
●		oddeleni	0.19%	0.07%	27.60%	0.00%	3.14%
●		mzda	2.49%	0.02%	48.78%	0.00%	0.45%

Zpracování: RapidMiner, vlastní (zdroj: Kaggle)

Jako nekvalitní, indikována červenou barvou, a tudíž nevhodná pro modelování je na základě kritérií CISMT vyhodnocena proměnná *povyseni*, a to na základě kritéria S, tj. stability, udávajícího hodnotu 97,87 %. Tzn. v případě této proměnné jsou její hodnoty z téměř 98 % shodné. Tudíž tato proměnná je téměř unární. Tato proměnná by vzhledem k této skutečnosti neměla být využita. Avšak v případě, pokud by v nadcházející fázi (fázi modelování) vykazovaly vytvořené predikční modely nízkou predikční sílu, může být tato proměnná do modelu opět zahrnuta a lze poté zkoumat, jak se změní predikční schopnost modelů.

Proměnná *hodnoceni* je vyhodnocena s varováním, jež je indikováno žlutou barvou. Je tomu tak z důvodu nízké korelace s cílovou proměnnou. Takováto proměnná pravděpodobně nepřispěje k vylepšení predikční síly modelů, avšak vzhledem k tomu, že se jedná pouze o varování a její ponechání by nemělo během procesu modelování působit problémy, v datech zůstane ponechána. Její významnost resp. nevýznamnost může být potvrzena či vyvrácena např. při sestavování rozhodovacího stromu, jenž umí vybírat významné proměnné během samotného modelování.

Rozdělení na data trénovací a testovací

V závěru fáze přípravy dat, byl datový soubor na základě poznatků Ayyadevara (2018) nabytých v teoretické části práce, v kapitole věnované hodnocení kvality modelu, rozdělen na data trénovací a data testovací v poměru 60:40.

4.4 Modelování

Během fáze modelování byly postupně vytvořeny čtyři klasifikační modely, kterými jsou naivní Bayesův klasifikátor, logistická regrese, rozhodovací strom a náhodný les. Každý z těchto modelů byl nejprve učen na datech trénovacích a poté byla otestována jeho predikční schopnost aplikací na data testovací. Úspěšnost predikce je vyhodnocena pomocí matic záměn a z ní vyplývajících hodnocení kvality modelu, kterými jsou celková přesnost, precisnost, specificita či sensitivita.

Z matice záměn je možné získat výpočtem různá hodnocení kvality modelu. Jedná se například o celkovou správnost či též přesnost (angl. accuracy), chybu (angl. error), senzitivitu (angl. sensitivity), specificitu (angl. specificity).

4.4.1 Naïve Bayes

Prvním testovaný modelem je Naïve Bayes, též naivní Bayesův klasifikátor.

Tato metoda je citlivá na chybějící hodnoty a proto je zapotřebí se vždy před modelováním ujistit, že žádné chybějící hodnoty data neobsahují a pokud obsahují, ošetřit je. Ve fázi přípravy dat bylo ověřeno, že soubor neobsahuje žádné chybějící hodnoty.

Výhodou metody Naïve Bayes je její jednoduchost, proto je nazývána naivní. Tato jednoduchost se projevuje i ve skutečnosti, že metoda Naïve Bayes nevyžaduje náročné ladění parametrů.

Metoda pracuje s kategorickými proměnnými. V případě číselných využívá průměrných hodnot, standardních odchylek a normálního rozdělení k odhadu pravděpodobností hodnot.

Za účelem předejití možnému zkreslení dat, v případě, kdy by se v trénovacích datech určitá hodnota některého prediktoru nevyskytovala a mohlo by se stát, že by podmíněná pravděpodobnost poté vyšla nulová, je pomocí operátoru *Naïve Bayes* v programu RapidMiner aplikována Laplaceova korekce.

Model Naïve Bayes byl v softwarovém nástroji RapidMiner nejprve učen na trénovacích datech a poté byl model aplikován na data testovací.

Predikční sílu modelu je možné analyzovat pomocí matice záměn (Tabulka 6), ve které je porovnáván počet záznamů správně a nesprávně klasifikovaných.

Tabulka 6: Matice záměn modelu Naïve Bayes

	true 0	true 1	class precision
pred. 0	3031	422	87.78%
pred. 1	234	598	71.88%
class recall	92.83%	58.63%	

Zpracování: RapidMiner, vlastní

Z hlediska přesnosti z matice záměn vyplývá, že hodnotu 1, tj. odchod zaměstnance, se podařilo metodě Naïve Bayes správně predikovat v necelých 72 % případů. V případě predikce hodnoty 0 byla úspěšná z necelých 88 %.

Dále je možné z matice záměn vyčíst, kolik procent hodnot dané proměnné se podařilo odhalit. V případě hodnoty 1, odchodu zaměstnance, se pomocí metody Naïve Bayes podařilo odhalit necelých 59 % odchodů z celkového počtu. V případě hodnoty 0 byl naivní Bayesův klasifikátor úspěšný z 93 %.

Z četností obsažených v matici záměn je možné vypočíst další míry hodnocení kvality modelu, jako jsou celková přesnost, klasifikační chyba, precisnost, sensitivita a specificita. Tyto hodnoty jsou uvedeny v následující tabulce (Tabulka 7).

Tabulka 7: Hodnocení kvality modelu Naïve Bayes

Criterion	Value
Accuracy	85.2%
Classification Error	14.8%
AUC	86.0%
Precision	71.9%
Recall	62.2%
Sensitivity	62.2%
Specificity	92.4%

Zpracování: RapidMiner, vlastní

Celková přesnost (angl. accuracy) predikce modelu Naïve Bayes je 85 %, tzn. 85 % ze všech hodnot predikovaných modelem při aplikaci na testovací data se shoduje se skutečným hodnotami a bylo tedy klasifikováno správně. Nesprávně bylo tudíž klasifikováno 15 % všech záznamů.

Precisnost (angl. precision) udává procento z predikovaných hodnot 1, tj. nastání odchodu zaměstnance, které byly hodnotami 1 i ve skutečnosti, tj. v testovacích datech. Model dosahuje precisnosti 72 %, tzn. 72 % z predikovaných hodnot 1, odpovídá i skutečným odchodům.

Úplnost, též sensitivita (angl. recall či sensitivity) udává, kolik ze skutečných hodnot 1, tj. odchodů zaměstnanců, se modelu podařilo nalézt. Hodnota úplnosti 59 % tudíž znamená, že 59 % odchodů zaměstnanců model odhalil a 41 % odchodů zaměstnanců neodhalil.

Specificita (angl. specificity) vyjadřuje, kolik hodnot 0, tj. nenastání odchodu zaměstnance, se modelu podařilo odhalit. V tomto případě byl Naïve Bayes úspěšný z 93 %.

Cílem je predikce odchodu zaměstnanců, avšak z uvedeného hodnocení modelu vyplývá, že byť je z 85 % přesný, z celkového počtu odchodů se mu podařilo odhalit pouze 59 % odchodů.

Za účelem zlepšení tohoto výsledku byl model optimalizován na základě poznatků Larose (2015) nabytých v teoretické části práce, v kapitole o přípravě dat, týkajících se diskretizace (binning či též binování) dat číselných prediktorů, která může v případě metod preferujících či vyžadujících prediktory kategorického typu, vylepšit jejich predikční sílu. Číselné prediktory spojitého typu, tj. *hodin_m*, *hodnoceni* a *spokojenost*, byly postupně binováním dat transformovány na prediktory kategorické, tj. rozděleny každá do určitého počtu binů nahrazených středními hodnotami reprezentujícími jednotlivé biny. Byly testovány různé kombinace počtu binů a metod jejich vytvoření u uvedených proměnných a z dosažených výsledků je zde prezentován model s nejlepším dosaženým výsledkem. Tím je model učený a testovaný na datech, v nichž byly spojité proměnné transformovány na kategorické pomocí diskretizace do 10 binů, přičemž tyto biny byly vytvořeny metodou založenou na shodném rozsahu intervalů jednotlivých binů.

Model byl následně opětovně učen a vyhodnocen na transformovaných trénovacích a testovacích datech. Hodnocení kvality optimalizovaného modelu vychází z následující matice záměn (Tabulka 8).

Tabulka 8: Matice záměn modelu Naïve Bayes po optimalizaci

	true 0	true 1	class precision
pred. 0	3207	345	90.29%
pred. 1	58	675	92.09%
class recall	98.22%	66.18%	

Zpracování: RapidMiner, vlastní

Optimalizací modelu využitím binování dat se podařilo zlepšit správnost predikce odchodu zaměstnance (hodnoty 1) ze 72 % na 92 %, tj. o 20 %. O 5 % úspěšnější je po optimalizaci i predikce hodnoty 0, tj. případu, kdy odchod zaměstnance nenastal. Z celkového počtu odchodů se modelu po optimalizaci podařilo odhalit 66 %, což je o 7 % více, než před optimalizací.

Následující tabulka (Tabulka 9) obsahuje hodnoty reprezentující mimo jiné celkovou přesnost, klasifikační chybu, precisnost, sensitivitu a specificitu optimalizovaného modelu.

Tabulka 9: Hodnocení kvality modelu Naïve Bayes po optimalizaci

Criterion	Value
Accuracy	90.6%
Classification Error	9.4%
AUC	94.9%
Precision	92.1%
Recall	66.2%
Sensitivity	66.2%
Specificity	98.2%

Zpracování: RapidMiner, vlastní

Celková přesnost modelu se zvýšila z 84,7 % na 90,6 %. Precisnost modelu, vyjadřující procento správně predikovaných odchodů, se též zvýšila ze 72 % na 92 %. Avšak úplnost resp. sensitivita, vyjadřující počet odchodů, které se optimalizovanému modelu podařilo odhalit, se zlepšila pouze o necelých 6 %, tj. dosáhla 66 %.

4.4.2 Logistická regrese

Druhým testovaným modelem je model binární logistické regrese.

Podmínkou aplikace tohoto modelu je binární cílová proměnná. Touto proměnnou je proměnná *odchod*, nabývající hodnoty 1 v případě nastání odchodu zaměstnance a v opačném případě nabývající hodnoty 0. Tato podmínka je tudíž splněna.

Vysvětlující proměnné neboli prediktory vstupující do modelu logistické regrese mohou být jak kategorické, tak číselné.

V případě kategorických proměnných možnost jejich zpracování logistickou regresí předpokládá jejich transformaci na dummy proměnné. Kategorickými proměnnými jsou *oddeleni*, *mzda* a *uraz*. Tyto proměnné byly tudíž ošetřeny v programu RapidMiner pomocí operátoru *Nominal to Numerical* a nastavení parametru kódování na hodnotu *Dummy coding*.

Model byl vytvořen pomocí operátoru *Logistic Regression* v programu RapidMiner. Nejprve byl učen na datech trénovacích a poté testován na datech testovacích.

Následující tabulka (Tabulka 10) obsahuje ve druhém sloupci (*Coefficient*) odhadnuté koeficienty logistické regrese. V případě číselných proměnných vyjadřují změnu šance nastání jevu při změně hodnoty dané proměnné o jednotku, za předpokladu, že hodnoty ostatních proměnných zůstanou nezměněné. V případě proměnných kategorických jednotlivé koeficienty vyjadřují změnu mezi danou kategorií a kategorií, jež byla použita jako referenční, za předpokladu, že hodnoty ostatních proměnných zůstanou nezměněné.

Tabulka 10: Model logistické regrese

Attribute ↑	Coefficient	Std. Coefficient	Std. Error	z-Value	p-Value
Intercept	0.175	-1.078	0.190	0.920	0.358
hodin_m	0.005	0.225	0.001	6.513	0.000
hodnoceni	0.691	0.119	0.199	3.475	0.001
mzda.2	-0.486	-0.486	0.062	-7.880	0.000
mzda.3	-1.991	-1.991	0.181	-11.007	0
oddeleni.administrativa	0.207	0.207	0.167	1.239	0.215
oddeleni.hr	0.276	0.276	0.165	1.672	0.094
oddeleni.management	-0.380	-0.380	0.204	-1.867	0.062
oddeleni.marketing	0.013	0.013	0.166	0.080	0.936
oddeleni.prodej	0.047	0.047	0.119	0.397	0.692
oddeleni.produktovy_mng	0.098	0.098	0.159	0.618	0.537
oddeleni.provoz	0.234	0.234	0.130	1.802	0.072
oddeleni.vyroba	0.255	0.255	0.124	2.048	0.041
oddeleni.vyzkum_vyvoj	-0.440	-0.440	0.183	-2.403	0.016
pocet_let	0.283	0.409	0.021	13.276	0
projekty	-0.311	-0.382	0.029	-10.826	0
spokojenost	-4.042	-1.004	0.132	-30.584	0
uraz.1	-1.506	-1.506	0.120	-12.543	0

Zpracování: RapidMiner, vlastní (zdroj: Kaggle)

Tabulka (Tabulka 10) rovněž obsahuje tzv. p-hodnoty (*p-value*), pomocí kterých lze posuzovat statistickou významnost jednotlivých proměnných pro model. Abbott (2014) jako typickou statistickou hladinu významnosti uvádí hodnotu 0,05 pro zařazení proměnné a hodnotu 0,1 pro vyřazení proměnné.

Všechny číselné proměnné, tj. *hodin_m*, *hodnoceni*, *pocet_let*, *projekty*, *spokojenost*, mají p-hodnotu nižší než je statistická hladina významnosti 0,05 a jsou tudíž pro model statisticky významné. Z kategorických proměnných se jako statisticky významné ukázaly kategorie *vyroba* a *vyzkum_vyvoj* proměnné *oddeleni*. Rovněž i proměnné *mzda* a *uraz*.

Predikční síla modelu logistické regrese je vyhodnocena pomocí matice záměn (Tabulka 11), porovnávající četnosti správně klasifikovaných a nesprávně klasifikovaných hodnot cílové proměnné *odchod*.

Tabulka 11: Matice záměn modelu logistické regrese

	true 0	true 1	class precision
pred. 0	3091	785	79.75%
pred. 1	174	235	57.46%
class recall	94.67%	23.04%	

Zpracování: RapidMiner, vlastní

Model logistické regrese správně klasifikoval 57 % záznamů s hodnotou 1, představujících odchod zaměstnance. V případě hodnoty 0, reprezentující nenastání odchodu, byl model úspěšný v 78 % případů.

23 % z celkového počtu hodnot proměnné 1, tj. odchodů, se modelu logistické regrese podařilo nalézt. V případě hodnoty 0 byl model úspěšný z 95 %.

Z četností uvedených v matici záměn byly pro model logistické regrese vypočteny kritéria hodnocení kvality celková přesnost, klasifikační chyba, precisnost sensitivita a specificita (Tabulka 12).

Tabulka 12: Hodnocení kvality modelu logistické regrese

Criterion	Value
Accuracy	77.6%
Classification Error	22.4%
AUC	79.1%
Precision	57.5%
Recall	23.0%
Sensitivity	23.0%
Specificity	94.7%

Zpracování: RapidMiner, vlastní

Celková přesnost (angl. accuracy) predikce modelu logistické regrese je 78 %, což znamená, že 78 % ze všech hodnot predikovaných modelem během aplikace na data testovací se shodovalo se skutečným hodnotami a bylo tak správně klasifikováno. Nesprávně bylo klasifikováno 22 % všech záznamů.

Precisnost (angl. precision) udává procento z predikovaných hodnot 1, tj. odchodů zaměstnanců, které byly hodnotami 1 i ve skutečnosti v testovacích datech. Model dosahuje precisnosti 57 %, tzn. 43 % z predikovaných hodnot 1, neodpovídá skutečným odchodům.

Úplnost, též sensitivita (angl. recall či sensitivity) udává, kolik ze skutečných hodnot 1, tj. odchodů zaměstnanců, se modelu podařilo nalézt. Hodnota úplnosti 23 % tudíž znamená, že 23 % odchodů zaměstnanců model odhalil a 77 % odchodů zaměstnanců neodhalil.

Specificita (angl. specificity) vyjadřuje, kolik hodnot 0, tj. nenastání odchodu zaměstnance, se modelu podařilo odhalit. V tomto případě byla logistická regrese úspěšná z 95 %.

Vytvořený model logistické regrese, vykazující celkovou přesnost 76 % a úplnost neboli sensitivitu, tj. počet odchodů, které se modelu podařilo odhalit, pouze 23 %, byl za účelem zlepšení výsledků predikce následně optimalizován. Optimalizace spočívala nejprve

v odstranění statisticky nevýznamné proměnné *oddeleni*, což však k zlepšení kvality modelu nepřispělo. Hodnoty celkové přesnosti a další míry se změnily pouze v rozsahu maximálně 0,5 %.

Úspěšnější se ukázal postup optimalizace spočívající v binningu dat, což, jak bylo zjištěno v rešerši v části věnované přípravě dat, je dle Larose (2015) jedna z možných metod vedoucích ke snížení chyby modelu. Číselné prediktory spojitého typu, tj. *hodin_m*, *hodnoceni* a *spokojenost*, byly postupně transformovány na kategorické prostřednictvím rozdělení na biny následně nahrazené příslušnými středními hodnotami binů. Testovány byly různé kombinace počtu binů a metod jejich vytvoření u uvedených proměnných, přičemž nejlepších výsledků bylo dosaženo v případě vytvoření 10 binů pro každou uvedenou proměnnou. Tyto biny byly vytvořeny metodou založenou na shodném rozsahu intervalů jednotlivých binů.

Následovalo opětovné učení modelu na trénovacích datech a poté jeho testování na datech validačních. Hodnocení kvality optimalizovaného modelu je založené na následující matici záměn (Tabulka 13).

Tabulka 13: Matice záměn modelu logistické regrese po optimalizaci

	true 0	true 1	class precision
pred. 0	3180	328	90.65%
pred. 1	85	692	89.06%
class recall	97.40%	67.84%	

Zpracování: RapidMiner, vlastní

V porovnání s původním modelem bylo správně klasifikováno 89 % odchodů zaměstnanců, což je zlepšení o 32 %. Též v případě správné klasifikace případů, kdy odchod nenastal, došlo ke zlepšení o 11 %.

Z celkového počtu odchodů zaměstnanců se modelu logistické regrese podařilo odhalit 68 % odchodů, což je zlepšení o 45 % oproti původnímu modelu.

Tabulka 14: Hodnocení kvality modelu logistické regrese po optimalizaci

Criterion	Value
Accuracy	90.4%
Classification Error	9.6%
AUC	94.8%
Precision	89.1%
Recall	67.8%
Sensitivity	67.8%
Specificity	97.4%

Zpracování: RapidMiner, vlastní

Tabulka 14 uvádí celkovou přesnost modelu logistické regrese, která se zvýšila o 12 %. V případě precisnosti vyjadřující procento správně klasifikovaných odchodů došlo ke zvýšení o 31 % oproti původnímu modelu. Úplnost neboli sensitivita, tj. kolik odchodů z celkového počtu, se modelu podařilo odhalit, se zvýšila o 45 %.

4.4.3 Rozhodovací strom

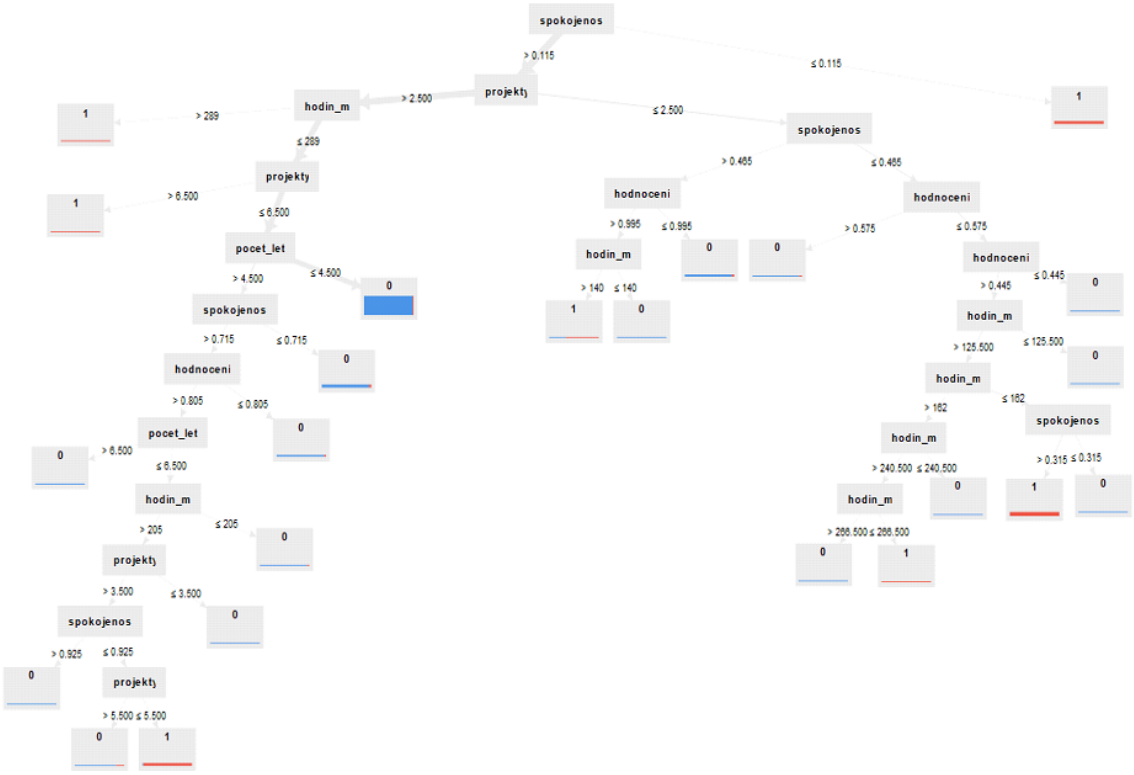
Třetím testovaným modelem je model rozhodovacího stromu.

Model je sestaven v programu RapidMiner pomocí operátoru *Model (Decision Tree)*.

Rozhodovací strom se skládá z kořene, který je pouze jeden, a dále se opakovaně dělí na základě podmínek v bodech, které se nazývají uzly. V případě, kdy se uzel již na základě žádné podmínky dále nerozděluje, je nazýván jako uzel terminální resp. list. V případě, kdy se uzly dělí vždy na dvě větve, jedná se o rozhodovací strom binární.

Následující graf zobrazuje vytvořený model rozhodovacího stromu.

Graf 3: Model rozhodovacího stromu



Zpracování: RapidMiner, vlastní (zdroj: Kaggle)

Rozhodovací strom představuje sadu pravidel generujících predikovanou hodnotu. Každé pravidlo je možné číst postupně od kořene stromu do vybraného terminálního uzlu resp. listu.

Výhodou rozhodovacích stromů je jejich jednoduchá aplikace, kvalita predikce a i snadná interpretace. Umí pracovat s numerickými i kategorickými proměnnými. Též dokáže vybírat podstatné prediktory.

V případě vytvořeného modelu nejsou obsaženy proměnné *oddeleni*, *uraz* a *mzda*. To znamená, že právě tyto proměnné byly modelem vyhodnoceny jako nevýznamné pro predikci, tj. jejich přidání by již nepřispělo k vylepšení predikce.

Kvalita modelu rozhodovacího stromu byla testována pomocí matice záměn, porovnávající četnosti správně klasifikovaných a nesprávně klasifikovaných hodnot cílové proměnné.

Tabulka 15: Matice záměn modelu rozhodovacího stromu

	true 0	true 1	class precision
pred. 0	3237	190	94.46%
pred. 1	28	830	96.74%
class recall	99.14%	81.37%	

Zpracování: RapidMiner, vlastní

Z hlediska přesnosti z matice záměn vyplývá, že hodnotu 1, tj. odchod zaměstnance, se podařilo rozhodovacímu stromu správně predikovat v necelých 97 % případů. V případě predikce hodnoty 0 byl úspěšný z 94 %.

Z celkového počtu odchodu zaměstnanců se rozhodovacímu stromu podařilo odhalit 81 %. Případy nenastání odchodu odhalil z 99 %.

Následně byl model rozhodovacího stromu z hlediska kvality posouzen na základě výpočtu vycházejících z matice záměn. Konkrétně byly vyhodnoceny celková přesnost, klasifikační chyba, precisnost a též sensitivita a specificita (Tabulka 16).

Tabulka 16: Hodnocení kvality modelu rozhodovacího stromu

Criterion	Value
Accuracy	94.9%
Classification Error	5.1%
AUC	97.1%
Precision	96.7%
Recall	81.4%
Sensitivity	81.4%
Specificity	99.1%

Zpracování: RapidMiner, vlastní

Celková přesnost (angl. accuracy) predikce rozhodovacího stromu je 95 %, což znamená, že pouze 5 % ze všech hodnot predikovaných modelem během aplikace na data testovací se neshodovalo se skutečným hodnotami.

Precisnost (angl. precision) vyjadřuje procento predikovaných odchodů zaměstnanců, které nastaly též ve skutečnosti, tj. shodují se hodnotami v testovacích datech. Rozhodovací strom byl v tomto případě úspěšný z 97 %.

Úplnost neboli sensitivita (angl. recall, sensitivity) udává, kolik procent ze skutečných odchodů zaměstnanců se modelu podařilo nalézt. Rozhodovací strom jich odhalil 81 %.

Specificita (angl. specificity) vyjadřující kolik nenastání odchodu zaměstnance se modelu podařilo odhalit, dosáhla v případě rozhodovacího stromu 95 %.

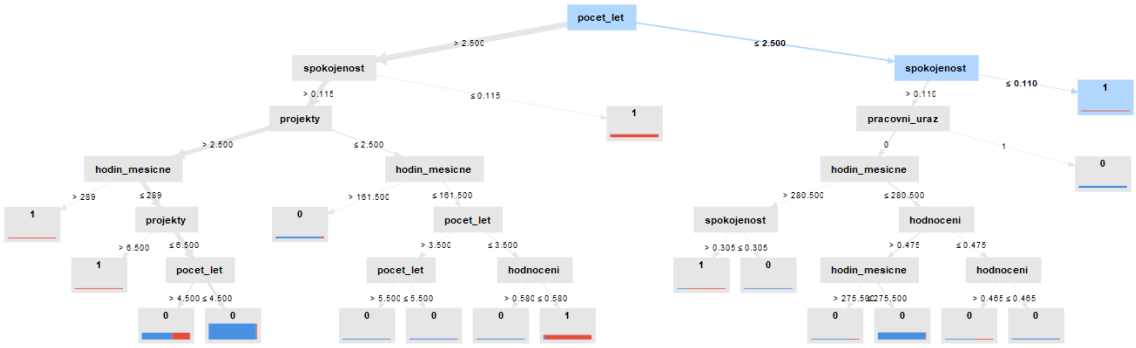
4.4.4 Náhodný les

Čtvrtým testovaným modelem je náhodný les.

Náhodný les spadá do kategorie tzv. ensemble metod, založených na vytvoření vícero modelů stejného typu a následném zkombinování jejich výsledků do konečné predikce.

Model je sestaven v programu RapidMiner pomocí operátoru *Model (Random Forest)*.

Tabulka 17: Ukázka jednoho z vytvořených stromů náhodného lesa



Zpracování: RapidMiner, vlastní (zdroj: Kaggle)

Kvalita modelu náhodného lesa byla posouzena na základě matice záměn (Tabulka 18), porovnávající četnosti správně a nesprávně klasifikovaných záznamů.

Tabulka 18: Matice záměn modelu náhodného lesa

	true 0	true 1	class precision
pred. 0	3181	145	95.64%
pred. 1	84	875	91.24%
class recall	97.43%	85.78%	

Zpracování: RapidMiner, vlastní

Z hlediska přesnosti z matice záměn vyplývá, že hodnotu 1, tj. odchod zaměstnance, se podařilo náhodnému lesu správně predikovat v necelých 91 % záznamů. V případě predikce hodnoty 0 byl úspěšný z 96 %. Z celkového počtu odchodu zaměstnanců náhodný les odhalil 87 %. Případy nenastání odchodu odhalil z 97 %.

Následně byl model rozhodovacího stromu z hlediska kvality posouzen pomocí celkové přesnosti, klasifikační chyby, precisnosti a též sensitivity a specificity (Tabulka 19).

Tabulka 19: Hodnocení kvality modelu náhodného lesa

Criterion	Value
Accuracy	94.7%
Classification Error	5.3%
AUC	97.7%
Precision	91.3%
Recall	85.8%
Sensitivity	85.8%
Specificity	97.4%

Zpracování: RapidMiner, vlastní

Celková přesnost (angl. accuracy) predikce náhodného lesa je 95 %. Precisnost (angl. precision) vyjadřující procento správně predikovaných odchodů zaměstnanců dosáhla 91 %. Úplnost či sensitivita (angl. recall, sensitivity) udávající, kolik procent ze skutečných odchodů zaměstnanců se modelu podařilo nalézt dosáhla 86 %.

Specificita (angl. specificity) vyjadřující kolik procent nenastání odchodu zaměstnance se modelu podařilo nalézt, dosáhla v případě náhodného lesa 97 %.

4.5 Vyhodnocení výsledků

Předposlední fází datové analýzy je fáze vyhodnocení výsledků, v které je porovnána predikční schopnost vytvořených modelů naivního Bayesova klasifikátoru, logistické regrese, rozhodovacího stromu a náhodného lesa, a na základě zjištěných poznatků je vybrán nejvhodnější model.

Následující tabulka (Tabulka 20) obsahuje přehled hodnocení vytvořených modelů.

Tabulka 20: Porovnání hodnocení klasifikačních modelů

Hodnocení klasifikátoru	Naïve Bayes	Logistická regrese	Rozhodovací strom	Náhodný les
Celková přesnost	91%	90%	95%	95%
Klasifikační chyba	9%	10%	5%	5%
AUC	95%	95%	97%	97%
Precisnost	92%	89%	97%	91%
Sensitivita, úplnost	66%	68%	81%	86%
Specificita	93%	97%	99%	97%

Zpracování: vlastní

Z hlediska celkové přesnosti modelů se jako nejvhodnější ukázal model rozhodovacího stromu a náhodného lesa, které oba dosáhly celkové přesnosti 95 %.

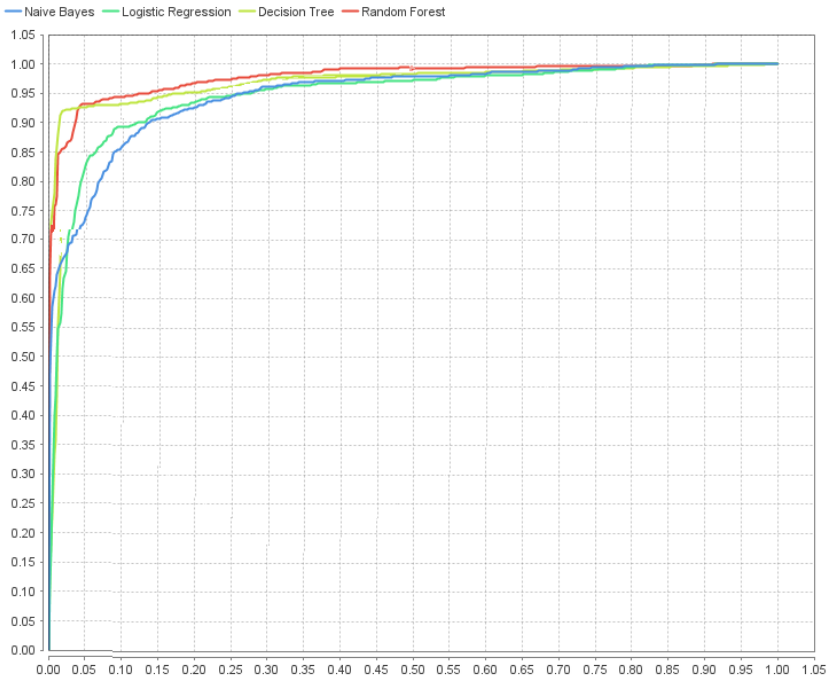
Z hlediska precisnosti, vyjadřující procento správně predikovaných odchodů zaměstnanců, dosáhl nejlepšího výsledku 97 % model rozhodovacího stromu.

Dle dosažené hodnoty sensitivity neboli úplnosti, představující procento odchodů zaměstnanců, které se podařilo danému modelu odhalit, nejlepšího výsledku 86 % dosáhl náhodný les.

V případě specificity, značící procento nenastání odchodu zaměstnance, které se podařilo modelu odhalit, nejlepšího výsledku 99 % dosáhl rozhodovací strom.

Na základě porovnání hodnot AUC (Area Under Curve) resp. ROC křivek znázorněných v následujícím grafu (Graf 4), dosáhly rozhodovací strom i náhodný les shodného nejlepšího výsledku 97 %, což je hodnota značící velmi dobrou predikci (Mandrekar, 2010).

Graf 4: Srovnání ROC křivek modelů

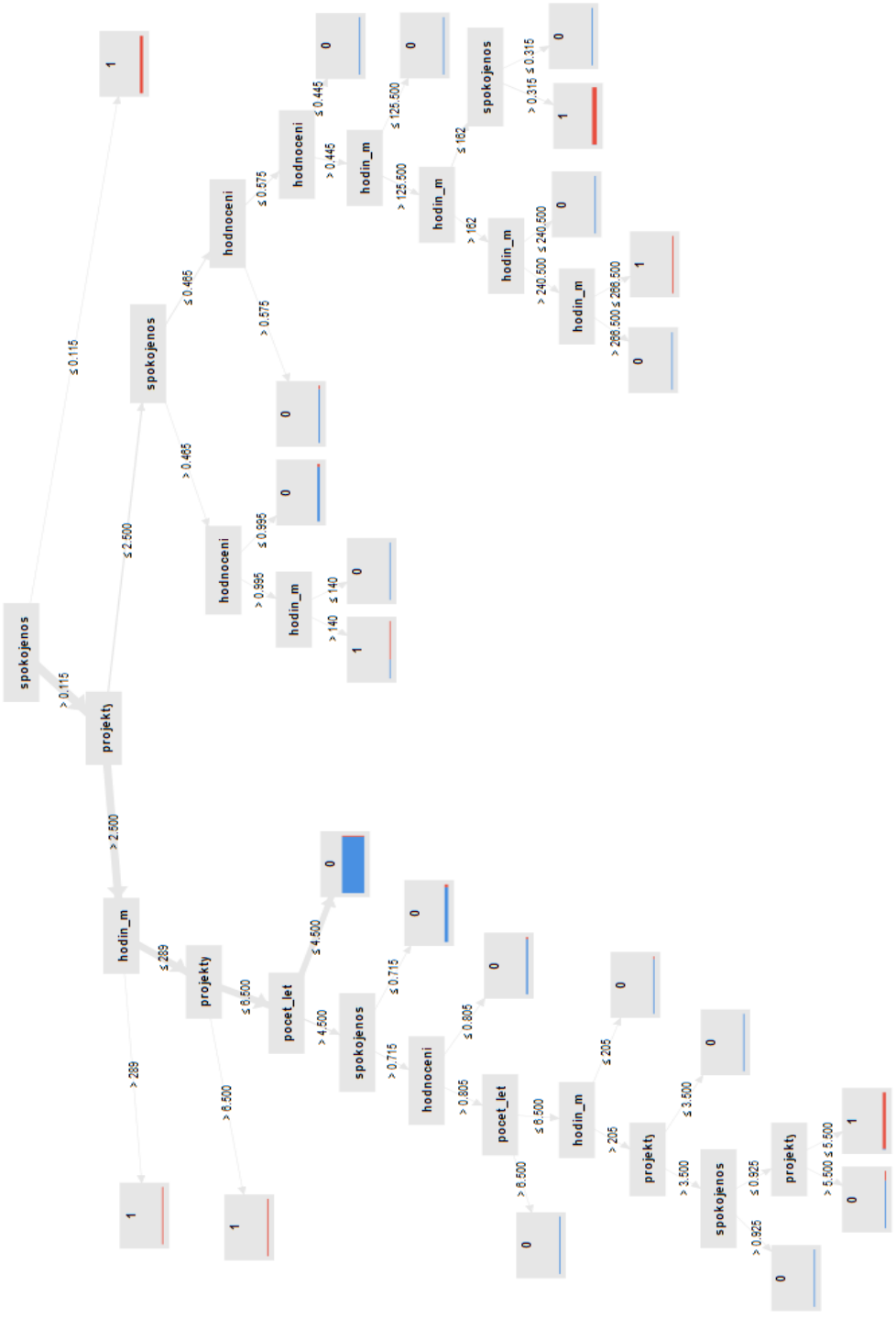


Zpracování: RapidMiner, vlastní

Vzhledem k srovnatelné kvalitě modelů rozhodovacího stromu a náhodného lesa, byl pro svou snadnou aplikaci, schopnost vybrat podstatné prediktory a též snadnou interpretaci a tvorbu rozhodovacích pravidel, vybrán jako nejvhodnější model rozhodovacího stromu.

Následující graf (Graf 5) zobrazuje výsledný model rozhodovacího stromu.

Graf 5: Model rozhodovacího stromu



Zpracování: RapidMiner, vlastní (zdroj: Kaggle)

Faktory ovlivňující odchod zaměstnance

Rozhodovací strom disponuje užitečnou vlastností spočívající ve schopnosti během své konstrukce odlišit prediktory podstatné od nepodstatných. Výsledný model rozhodovacího stromu neobsahuje proměnné *oddeleni* a *uraz* a *mzda*, z čehož vyplývá, že tyto proměnné byly vyhodnoceny v tomto případě jako nevýznamné pro predikci odchodu zaměstnance pomocí rozhodovacího stromu.

Program RapidMiner nabízí též nástroj hodnotící významnost prediktorů. Tu vyhodnocuje pomocí vah vyjadřujících sumu vylepšení predikce na uzlech stromu dosaženého zahrnutím prediktoru (Mierswa, Klinkenberg 2020). Jako tři nejvýznamnější prediktory byly vyhodnoceny proměnné *projekty*, *spokojenost* a *hodin_m*.

Tabulka 21: Významnost prediktorů

Attribute	Weight
projekty	0.108
spokojenost	0.080
hodin_m	0.054
hodnoceni	0.043
pocet_let	0.027

Zpracování: RapidMiner, vlastní (zdroj: Kaggle)

Na základě srovnání významnosti prediktorů s poznatky Srivastava a Tiwari (2020), kteří jako faktory nejvíce ovlivňující odchody zaměstnanců vyhodnotili časté pracovní cesty, časté střídání zaměstnání, nutnost přesčasů, spokojenost zaměstnance a angažovanost v zaměstnání, je možné potvrdit, že počet projektů, do nichž byl zaměstnanec zapojen, je významným faktorem. Počet projektů spadá do kategorie faktorů vyjadřujících angažovanost v zaměstnání. Rovněž v případě spokojenosti zaměstnance lze konstatovat, že byl též potvrzen jako významný faktor ovlivňující odchody zaměstnanců. Podobně prediktor obsahující počet hodin měsíčně, byl potvrzen jako významný faktor vzhledem k tomu, že vysoký počet odpracovaných hodin v měsíci odpovídá nutnosti vykonávat přesčasovou práci.

Závěrem fáze vyhodnocení výsledků je posouzení, zda vytvořený model splňuje cíle stanovené v prvé fázi. Vzhledem ke skutečnosti, že jako nejvhodnější model predikující odchod zaměstnanců byl vybrán a vytvořen rozhodovací strom a následně byly identifikovány faktory mající největší vliv na odchod zaměstnanců, lze konstatovat, že cíle byly splněny.

4.6 Využití výsledků

Poslední fáze analýzy dat dle metodiky CRISP-DM spočívá ve využití výsledků.

Kromě nasazení samotného vytvořeného modelu rozhodovacího stromu do praxe je také možné odvozovat z něj rozhodovací pravidla generující predikovanou hodnotu. Lze tak učinit postupným průchodem rozhodovacím stromem od jeho kořene do vybraného terminálního uzlu neboli listu. Nejvýznamnější pravidla vytvořená na základě výsledného rozhodovacího stromu jsou uvedena v následující tabulce (Tabulka 22).

Tabulka 22: Pravidla odvozená z rozhodovacího stromu

Pravidlo	Hodin měsíčně	Hodnocení	Počet let	Počet projektů	Spokojenost	Predikce
1	-	-	-	-	méně než 11,5 %	1
2	126 až 162	45 % až 58 %	-	méně než 3	32 % až 47 %	1
3	206 až 289	více než 81 %	4,5 až 6,5	4 až 6	72 % až 93 %	1
4	méně než 289	-	méně než 4,5	3 až 7	více než 11,5 %	0
5	méně než 289	-	více než 4,5	3 až 7	12 % až 72 %	0
6	-	47 % až 99,5 %	-	méně než 3	více než 11,5 %	0

Zpracování: vlastní

Nejjednodušším a zároveň silným pravidlem, s významným zastoupením, které lze z rozhodovacího stromu odvodit, je pravidlo číslo 1 reprezentované vyhodnocením pouhé jediné proměnné, kterou je proměnná *spokojenost*, jež v případě, že je menší nebo rovna hodnotě 0,115, tj. je menší než 11,5 % je pro daného zaměstnance rovnou predikována hodnota 1 indikující jeho odchod.

Pravidlo č. 2 rovněž predikuje odchod zaměstnance. Toto pravidlo spočívá ve vyhodnocení čtyř prediktorů, kterými jsou *hodin_m*, *hodnoceni*, *projekty* a *spokojenost*. Konkrétně se

jedná o zaměstnance, o nichž je známo, že průměrný počet odpracovaných hodin je v intervalu od 126 do 162. Během hodnocení zaměstnanců dosáhli skóre mezi 45 % a 58 %. Počet projektů, do nichž byli zapojeni, je menší než 3. A hodnocení jejich spokojenosti se nachází mezi 32 % a 47 %.

Dalším pravidlem predikujícím odchod zaměstnance je pravidlo číslo 3. Sestává z vyhodnocení hodnot prediktorů *hodin_m*, *hodnoceni*, *pocet_let*, *projekty* a *spokojenost*. Průměrný počet odpracovaných hodin měsíčně se nachází v intervalu od 206 do 289. Zaměstnanec dosáhl hodnocení přesahujícího 81 %. Délka zaměstnání je od 4,5 do 6,5 let. Počet projektů, do nichž byl zapojen je od 3 do 7. Spokojenost je vyhodnocena mezi 12 % a 72 %.

Z rozhodovacího stromu je možné odvozovat pravidla pro nenastání odchodu zaměstnance. Těmi jsou i pravidla číslo 4, 5 a 6. Například v případě pravidla číslo 4, bude-li určitý záznam o zaměstnanci obsahovat hodnoty prediktorů takové, že průměrný měsíční počet odpracovaných hodin bude menší než 289, délka zaměstnání bude více než 4,5 let, počet projektů bude od 3 do 7 a hodnocení spokojenosti zaměstnance bude přesahovat 11,5 %, bude takovémuto zaměstnanci predikována hodnota 0 značící jeho setrvání, tj. že odchod nenastane.

Takováto pravidla je možné využít jednak pro jednoduché vyhodnocování hrozících odchodů zaměstnanců a rovněž jako podklad pro analýzu možných preventivních opatření vedoucích k snížení počtu odchodů zaměstnanců.

5 Závěr

Cílem diplomové práce bylo vytvoření vhodného modelu predikujícího odchod zaměstnanců z organizace a rovněž identifikace faktorů tyto odchody nejvíce ovlivňujících.

Za účelem splnění cíle se práce v prvé, teoretické části, zabývala nejprve daty obecně a poté více tématu big data. Následně se věnovala analýze dat, přičemž se zaměřila na prediktivní modelování. Pozornost byla věnována též postupu zpracování prediktivní analýzy, konkrétně dle metodiky CRISP-DM. Vzhledem k povaze úlohy řešené v praktické části práce, kterou byla predikce cílové proměnné binárního typu, se teoretická část práce dále zaměřila na modely klasifikační, konkrétně na Naïve Bayes, logistickou regresi, rozhodovací strom a náhodný les. Pozornost byla věnována též nástrojům sloužícím k hodnocení kvality modelů a rovněž tématu lidských zdrojů, jakožto oblasti, v níž se problematika, řešená v praktické části, nachází.

Na část teoretickou navázala praktická část práce. Jejím cílem bylo vytvoření vhodného modelu predikujícího odchod zaměstnanců z organizace a následně pomocí vytvořeného modelu nalezení faktorů odchody zaměstnanců nejvíce ovlivňujících.

Analyzovaná data obsahovala 15000 plně anonymizovaných záznamů o zaměstnancích. Postup prediktivního modelování vycházel z metodiky CRISP-DM, jež sestává z šesti fázi. Těmito fázemi jsou porozumění problematice, porozumění datům, příprava dat, modelování, vyhodnocení a využití výsledků. Po seznámení se s problematikou a stanovením dílčích cílů prediktivní analýzy, byla pozornost věnována přípravě dat. Během přípravy dat byly vysvětlující proměnné transformovány tak, aby jejich typ odpovídal jejich obsahu. Rovněž byla prověřena kvalita proměnných. Poté byl soubor dat rozdělen na data trénovací a testovací. Následně byly vytvořeny modely Naïve Bayes, logistické regrese, rozhodovacího stromu a náhodného lesa. Dalším krokem bylo srovnání predikční síly vytvořených modelů. K tomuto účelu byly využity matice záměn jednotlivých modelů a z nich vycházející hodnocení, kterými byly celková přesnost,

precisnost, sensitivita a specificita. Dalšími nástroji, které byly využity k porovnání predikční schopnosti modelů, byly ROC křivky a hodnoty AUC vytvořených modelů. Srovnatelných nejlepších výsledků dosáhly modely rozhodovacího stromu a náhodného lesa. Jako nejvhodnější byl vybrán model rozhodovacího stromu, vzhledem ke svým vlastnostem, spočívajícím především ve snazší interpretaci a tvorbě rozhodovacích pravidel.

Jako faktory nejvíce ovlivňující odchody zaměstnanců byly identifikovány proměnné vypovídající o počtu projektů, do nichž byl zaměstnanec během svého působení u organizace zapojen, dále spokojenost zaměstnance a též průměrný měsíční počet odpracovaných hodin. Tyto nalezené nejvýznamnější faktory jsou v souladu s poznatky teoretické části práce, konkrétně kapitolou věnovanou odchodům zaměstnanců, kde Srivastava a Tiwari (2020) mezi faktory nejvíce ovlivňující odchody zaměstnanců uvádí nutnost přesčasů, spokojenost zaměstnance a angažovanost v zaměstnání.

Na základě znalosti faktorů nejvíce ovlivňujících odchod zaměstnanců, modelu rozhodovacího stromu a z něj odvozených pravidel je možné odhalovat hrozící odchody zaměstnanců. Rovněž mohou být využity jako podklad pro analýzu možných preventivních opatření vedoucích k snížení počtu odchodů zaměstnanců.

6 Seznam použitých zdrojů

ABBOTT, Dean, 2014. Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst. Vyd. 1. Wiley. New York. ISBN 9781118727935.

ARMSTRONG, Michael, 2007. Řízení lidských zdrojů. Vyd. 10. Praha: Grada Publishing a.s. ISBN: 978-80-247-1407-3.

AYYADEVARA, V. K. Pro Machine Learning Algorithms: A Hands-On Approach to Implementing Algorithms in Python and R. United States, CA: Apress, 2018. ISBN 978-1-4842-3564-5.

BERKA Petr, 2003. Dobývání znalostí z databází. Vyd. 1. Praha: Academia. ISBN 90-8020010629.

BEYER, Mark a LANEY, Douglas, 2012. The Importance of Big Data: A Definition [online]. Stanford: Gartner, 21.6. [cit. 16.6.2020] Dostupné z: www.gartner.com /doc /2057415/importance-big-data-definition

BUDÍNOVÁ Marie, 2017. Binární logistická regrese [online]. Brno: Ústav matematiky a statistiky, Masarykova univerzita. [cit. 16.10.2020]. Dostupné z: https://is.muni.cz/el/sci/jaro2017/MAS02/um/67880443/

COX, Michael a ELLSWORTH, David, 1997. Application-controlled demand paging for out-of-core visualization. In: Proceedings of the 8th conference on Visualization'97. USA: IEEE Computer Society Press Los Alamitos. ISBN: 1-58113-011-2.

DEAN, Jared, 2014. Big Data, Data Mining and Machine Learning: Value Creation for Business Leaders and Practitioners. Vyd. 1. New York: Wiley. ISBN: 9781118618042.

HOLČÍK, Jiří a Martin KOMENDA, 2015. Matematická biologie: e-learningová učebnice [online]. 1. vyd. Brno: Masarykova univerzita, 2015. [cit. 5.9.2020]. ISBN 978-80-210-8095-9.

HOLUBOVÁ, Irena, KOSEK, Jiří, MINAŘÍK, Karel a NOVÁK, David, 2015. Big Data a NoSQL databáze. Vyd. 1. Praha: Grada. ISBN: 978-8024754666.

INMON, William H., 2002. Building the data warehouse. Vyd. 3. New York: Wiley. ISBN 0471081302.

KAGGLE, 2020 [online]. © 2020 Kaggle Inc. San Francisco, California, US. [cit. 10.10.2020] Dostupné z: https://www.kaggle.com/

KLOUDA, Karel a VAŠATA, Daniel, 2020. Vytěžování znalostí z dat [online]. Praha: FIT ČVUT. [cit. 15.9.2020] Dostupné z: https://kam.fit.cvut.cz/bi-vzd/lectures/index.html

KOHAVI, R., BECKER., B, SOMMERFELD, D., 1997. Improving simple Bayes [online]. Poster Papers of the 9th Europen Conf. on Machine Learning. [cit. 15.9.2020] Dostupné z: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.26.7477

KOMPRDOVÁ, Klára, 2012. Rozhodovací stromy a lesy. Vyd. 1. Brno: Akademické nakladatelství Cerm. ISBN 978-80-7204-785-7.

KOTU Vitay, DESHPANDE Bala, 2019. Data Science: Concepts and Practice. Vyd. 2. Cambridge: Elsevier Inc. ISBN 978-0-12-814761-0.

LANEY, Doug, 2001. 3D Data Management: Controlling Data Volume, Velocity and Variety [online] Stamford: META Group Inc. 6.2. [cit. 16.6.2020]. Dostupné z: https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-ManagementControlling-Data-Volume-Velocity-and-Variety.pdf

LAROSE, Daniel T., LAROSE, Chantal D., 2015. Data mining and predictive analytics. Vyd. 2. New Jersey: Wiley & Sons, Inc. ISBN 978-1-118-11619-7.

LITSCHMANNOVÁ, Martina, 2011. Průzkumová analýza jednorozměrných dat (Teorie) [online] VŠB-TU Ostrava. [cit. 20.10.2020]. Dostupné z: http://gisak.vsb.cz/pan/source/data/skolenia/eda/Pruzkumova_analyza_dat.pdf

MANDREKAR, J. N., 2010. Receiver Operating Characteristic Curve in Diagnostic Test Assessment [online]. Biostatistics for Clinicians, Journal of Thoracic Oncology,

5 (9), 1315-1316. [cit. 15.12.2020]. Dostupné z: https://www.jto.org/article/S1556-0864(15)30604-3/fulltext

MANUABA I.B.K., INDRAJANI S., BAHANA R, 2020. The Evaluation of Superfised Classifier Models to Develop a Machine Learning API for Predicting Cardio-vascular Disease Risk [online]. ICIC Express Letters, 14 (3), 219-226. [cit. 15.12.2020]. ISSN: 1881-803X. Dostupné z: http://www.icicel.org/ell/contents/2020/3/el-14-03-02.pdf

MAYER-SCHONBERGER, Viktor, CUKIER, Kenneth, 2013. Big Data. Vyd. 1. Brno: Computer Press. ISBN 978-80-251-4119-9.

MELOUN, Milan a MILITKÝ, Jiří, 2012. Kompendium statistického zpracování dat. Vyd. 3. Praha: Karolinum. ISBN 978-8024621968.

MIERSWA, I., KLINKENBERG, R., 2020. RapidMiner Studio Educational 9.8.001 (rev: c9c089, platform: WIN64). RapidMiner GmbH. Dostupné z: https://rapidminer.com/educational-program/

NORMANDEAU, Kevin, 2013. Beyond Volume, Variety and Velocity is the Issue of Big Data Veracity. In: InsideBIGDATA.com [online]. 12.9. [cit. 17.6.2020]. Dostupné z: https://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-bigdata-veracity/

ROEBUCK, Kevin, 2011. Predictive Analysis: High-impact Emerging Technology - What You Need to Know: Definitions, Adoptions, Impact, Benefits, Maturity, Vendors. Vyd. 1. Ickenham: Tebbo. ISBN 1743042884.

SRIVASTAVA, Devesh Kumar, TIWARI, Pradeep Kumar, 2020. An analysis report to reduce the employee attrition within organizations [online]. In: Journal of Discrete Mathematical Sciences and Cryptography. [cit. 30.11.2020]. Vol. 23, No. 2, pp. 337-348, ISSN 2169-0065, DOI: 10.1080/09720529.2020.1721874 Dostupné z: https://doi.org/10.1080/09720529.2020.1721874